Package: tlda (via r-universe)

June 6, 2026

Title Tools for Language Data Analysis

Version 0.1.0.9000

Description Support functions and datasets to facilitate the analysis of linguistic data. The current focus is on the calculation of corpus-linguistic dispersion measures as described in Gries (2021) <doi:10.1007/978-3-030-46216-1_5> and Soenning (2025) <doi:10.3366/cor.2025.0326>. The most commonly used parts-based indices are implemented, including different formulas and modifications that are found in the literature, with the additional option to obtain frequency-adjusted scores. Dispersion scores can be computed based on individual count variables or a term-document matrix.

License MIT + file LICENSE

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Suggests boot, ggplot2, knitr, scales, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

Depends R (>= 3.5)

LazyData true

URL https://github.com/lsoenning/tlda

BugReports https://github.com/lsoenning/tlda/issues

VignetteBuilder knitr

Collate 'add_sampling_weights.R' 'biber150_brown.R' 'biber150_brown_genre.R' 'biber150_brown_macro_genre.R' 'biber150_ice_gb.R' 'biber150_ice_gb_genre.R' 'biber150_ice_gb_macro_genre.R' 'biber150_spokenBNC1994.R' 'biber150_spokenBNC2014.R' 'disp.R' 'disp_DA.R' 'disp_DKL.R' 'disp_DMB.R' 'disp_DP.R' 'disp_R.R' 'disp_S.R' 'dispersion_min_max_functions.R' 'metadata_brown.R' 'metadata_ice_gb.R' 'metadata_spokenBNC1994.R'

'metadata_spokenBNC2014.R' 'scale_folded_power.R'
'scale_y_dispersion.R'

Repository https://lsoenning.r-universe.dev

Date/Publication 2025-12-08 15:42:29 UTC

RemoteUrl https://github.com/lsoenning/tlda

RemoteRef HEAD

RemoteSha bad3a6dfedc17aaea43c5582b80488fd15faba6b

Contents

`add_sampling_weights` *Add column with sampling weights to a data frame*

Description

This function adds a new column `sampling_weights` to a data frame. The purpose of sampling weights is to adjust for a mismatch between sample and target population with regard to the distribution of one or more categorical variables. The population distribution can be specified, and sampling weights are then calculated to work as adjustment factors. The default is to assume a balanced population distribution (all levels represented equally), but custom population distributions can be specified.

Usage

```
add_sampling_weights(data, variable, population_distribution = NULL)
```

Arguments

`data`	A data frame
`variable`	Character string indicating the variable for whose levels sampling weights should be calculated
`population_distribution`	List (or data frame) specifying the population distribution of the levels; default is `NULL`, meaning that all levels are assumed to be represented equally

Details

This function takes as input a data frame, where observations (rows) represent a sample from a population. If the distribution of a categorical variable in the sample does not match the (known or assumed) distribution in the population, the function calculated sampling weights for the observations (rows). These account for the mismatch by up-weighting rows of those level that are underrepresented in the sample (relative to the population) and down-weighting rows belonging to levels that are overrepresented in the sample (relative to the population). If no population distribution is specified, all levels are assumed to be represented equally in the target population. Sampling weights are calculated on the basis of (i) the observed distribution of the variable in the sample, and (ii) the population distribution. For instance, if a specific subgroup (i.e. level) has a share of 10% in the sample, compared to 20% in the population, the sampling weight is 2.0 (20% divided by 10%). Sampling weights above 1 indicate up-weighting, sampling weights below 1 indicate down-weighting. The function prints out information on the sample and population distribution and the resulting weights.

Value

A data frame

Author(s)

Lukas Soenning

Examples

```
add_sampling_weights(
  data = metadata_ice_gb,
  variable = "mode")
```

`biber150_brown` *Distribution of Biber et al.'s (2016) 150 lexical items in the Brown Corpus (term-document matrix), using text files as corpus parts*

Description

This dataset contains text-level frequencies for the Brown Corpus (Francis & Kučera 1979) for a set of 150 word forms. The list of items was compiled by Biber et al. (2016) for methodological purposes, that is, to study the behavior of dispersion measures in different distributional settings. The items are intended to cover a broad range of frequency and dispersion levels.

Usage

```
biber150_brown
```

Format

`biber150_brown`**:**

A matrix with 151 rows and 500 columns

rows Length of text (`word_count`), followed by set of 150 items in alphabetical order (*a, able, ..., you, your*)

columns 500 texts, ordered by file name (e.g. "A01", "A02", ... , "R08", "R09")

Details

While Biber et al. (2016: 446) used 153 target items, the 150 word forms included in the present data set correspond to the slightly narrower selection of forms used in Burch et al. (2017: 214-216). These 150 word forms are listed next, in alphabetical order:

a, able, actually, after, against, ah, aha, all, among, an, and, another, anybody, at, aye, be, became, been, began, bet, between, bloke, both, bringing, brought, but, charles, claimed, cor, corp, cos, da, day, decided, did, do, doo, during, each, economic, eh, eighty, england, er, erm, etcetera, everybody, fall, fig, for, forty, found, from, full, get, government, ha, had, has, have, having, held, hello, himself, hm, however, hundred, i, ibm, if, important, in, inc, including, international, into, it,

just, know, large, later, latter, let, life, ltd, made, may, methods, mhm, minus, mm, most, mr, mum, new, nineteen, ninety, nodded, nought, oh, okay, on, ooh, out, pence, percent, political, presence, provides, put, really, reckon, say, seemed, seriously, sixty, smiled, so, social, somebody, system, take, talking, than, the, they, thing, think, thirteen, though, thus, time, tt, tv, twenty, uk, under, urgh, us, usa, wants, was, we, who, with, world, yeah, yes, you, your

The data are provided in the form of a term-document matrix, where rows denote the 150 items and columns denote the 500 texts in the corpus. Seven items do not occur in Brown (*aha, cor, cos, ltd, mhm, nought, pence)*). These are included in the term-document matrix with frequencies of 0 for all texts. Further, seven items are spelled differently in Brown (compared to the BNC, on which Biber et al.'s (2016) study is based): "u.s.a." (Brown) instead of "usa" (BNC), "inc." instead of "inc", "mr." instead of "mr", "ugh" instead of "urgh", "uh" instead of "er", "um" instead of "erm", and "hmm" instead of "hm".

The first row of the term-document matrix gives the length of the text (i.e. number of word and nonword tokens).

Source

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Francis, W. Nelson & Henry Kučera. 1979. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers (Brown)*. Providence, RI: Brown University.

`biber150_brown_genre` *Distribution of Biber et al.'s (2016) 150 lexical items in the Brown Corpus (term-document matrix), using genres as corpus parts*

Description

This dataset contains text-level frequencies for the Brown Corpus (Francis & Kučera 1979) for a set of 150 word forms. The list of items was compiled by Biber et al. (2016) for methodological purposes, that is, to study the behavior of dispersion measures in different distributional settings. The items are intended to cover a broad range of frequency and dispersion levels.

Usage

```
biber150_brown_genre
```

Format

`biber150_brown_genre`**:**

A matrix with 151 rows and 15 columns

rows Size of the genre (`word_count`), followed by set of 150 items in alphabetical order (*a, able, ..., you, your*)

columns 15 genres, ordered based on the sampling frame ("press_reportage", "press_editorial", ... ,"romance_love_story", "humour")

Details

While Biber et al. (2016: 446) used 153 target items, the 150 word forms included in the present data set correspond to the slightly narrower selection of forms used in Burch et al. (2017: 214-216). These 150 word forms are listed next, in alphabetical order:

a, able, actually, after, against, ah, aha, all, among, an, and, another, anybody, at, aye, be, became, been, began, bet, between, bloke, both, bringing, brought, but, charles, claimed, cor, corp, cos, da, day, decided, did, do, doo, during, each, economic, eh, eighty, england, er, erm, etcetera, everybody, fall, fig, for, forty, found, from, full, get, government, ha, had, has, have, having, held, hello, himself, hm, however, hundred, i, ibm, if, important, in, inc, including, international, into, it, just, know, large, later, latter, let, life, ltd, made, may, methods, mhm, minus, mm, most, mr, mum, new, nineteen, ninety, nodded, nought, oh, okay, on, ooh, out, pence, percent, political, presence, provides, put, really, reckon, say, seemed, seriously, sixty, smiled, so, social, somebody, system, take, talking, than, the, they, thing, think, thirteen, though, thus, time, tt, tv, twenty, uk, under, urgh, us, usa, wants, was, we, who, with, world, yeah, yes, you, your

The data are provided in the form of a term-document matrix, where rows denote the 150 items and columns denote the 15 genres in the corpus. Seven items do not occur in Brown (*aha, cor, cos, ltd, mhm, nought, pence)*). These are included in the term-document matrix with frequencies of 0 for all texts. Further, seven items are spelled differently in Brown (compared to the BNC, on which Biber et al.'s (2016) study is based): "u.s.a." (Brown) instead of "usa" (BNC), "inc." instead of "inc", "mr." instead of "mr", "ugh" instead of "urgh", "uh" instead of "er", "um" instead of "erm", and "hmm" instead of "hm".

The first row of the term-document matrix gives the size of the genre (i.e. number of word and nonword tokens).

Source

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Francis, W. Nelson & Henry Kučera. 1979. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers (Brown)*. Providence, RI: Brown University.

`biber150_brown_macro_genre`

Distribution of Biber et al.'s (2016) 150 lexical items in the Brown Corpus (term-document matrix), using macro-genres as corpus parts

Description

This dataset contains text-level frequencies for the Brown Corpus (Francis & Kučera 1979) for a set of 150 word forms. The list of items was compiled by Biber et al. (2016) for methodological purposes, that is, to study the behavior of dispersion measures in different distributional settings. The items are intended to cover a broad range of frequency and dispersion levels.

Usage

```
biber150_brown_macro_genre
```

Format

`biber150_brown_macro_genre`**:**

A matrix with 151 rows and 4 columns

rows Size of the macro genre (`word_count`), followed by set of 150 items in alphabetical order (*a, able, ..., you, your*)

columns 4 macro genres, ordered based on the sampling frame ("press", "general_prose", "learned", "fiction")

Details

While Biber et al. (2016: 446) used 153 target items, the 150 word forms included in the present data set correspond to the slightly narrower selection of forms used in Burch et al. (2017: 214-216). These 150 word forms are listed next, in alphabetical order:

a, able, actually, after, against, ah, aha, all, among, an, and, another, anybody, at, aye, be, became, been, began, bet, between, bloke, both, bringing, brought, but, charles, claimed, cor, corp, cos, da, day, decided, did, do, doo, during, each, economic, eh, eighty, england, er, erm, etcetera, everybody, fall, fig, for, forty, found, from, full, get, government, ha, had, has, have, having, held, hello, himself, hm, however, hundred, i, ibm, if, important, in, inc, including, international, into, it, just, know, large, later, latter, let, life, ltd, made, may, methods, mhm, minus, mm, most, mr, mum, new, nineteen, ninety, nodded, nought, oh, okay, on, ooh, out, pence, percent, political, presence, provides, put, really, reckon, say, seemed, seriously, sixty, smiled, so, social, somebody, system, take, talking, than, the, they, thing, think, thirteen, though, thus, time, tt, tv, twenty, uk, under, urgh, us, usa, wants, was, we, who, with, world, yeah, yes, you, your

The data are provided in the form of a term-document matrix, where rows denote the 150 items and columns denote the 4 macro genres in the corpus. Seven items do not occur in Brown (*aha, cor, cos, ltd, mhm, nought, pence)*). These are included in the term-document matrix with frequencies of 0 for all texts. Further, seven items are spelled differently in Brown (compared to the BNC, on which Biber et al.'s (2016) study is based): "u.s.a." (Brown) instead of "usa" (BNC), "inc." instead

of "inc", "mr." instead of "mr", "ugh" instead of "urgh", "uh" instead of "er", "um" instead of "erm", and "hmm" instead of "hm".

The first row of the term-document matrix gives the size of the genre (i.e. number of word and nonword tokens).

Source

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Francis, W. Nelson & Henry Kučera. 1979. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers (Brown)*. Providence, RI: Brown University.

`biber150_ice_gb` *Distribution of Biber et al.'s (2016) 150 lexical items in ICE-GB (term-document matrix), using text files as corpus parts*

Description

This dataset contains text-level frequencies for ICE-GB (Nelson et al. 2002) for a set of 150 word forms. The list of items was compiled by Biber et al. (2016) for methodological purposes, that is, to study the behavior of dispersion measures in different distributional settings. The items are intended to cover a broad range of frequency and dispersion levels.

Usage

```
biber150_ice_gb
```

Format

`biber150_ice_gb`**:**

A matrix with 151 rows and 500 columns

rows Length of text (`word_count`), followed by set of 150 items in alphabetical order (*a, able, ..., you, your*)

columns 500 texts, ordered by file name ("s1a-001","s1a-002", ... , "w2f-019", "w2f-020"))

Details

While Biber et al. (2016: 446) used 153 target items, the 150 word forms included in the present data set correspond to the slightly narrower selection of forms used in Burch et al. (2017: 214-216). These 150 word forms are listed next, in alphabetical order:

a, able, actually, after, against, ah, aha, all, among, an, and, another, anybody, at, aye, be, became, been, began, bet, between, bloke, both, bringing, brought, but, charles, claimed, cor, corp, cos,

da, day, decided, did, do, doo, during, each, economic, eh, eighty, england, er, erm, etcetera, everybody, fall, fig, for, forty, found, from, full, get, government, ha, had, has, have, having, held, hello, himself, hm, however, hundred, i, ibm, if, important, in, inc, including, international, into, it, just, know, large, later, latter, let, life, ltd, made, may, methods, mhm, minus, mm, most, mr, mum, new, nineteen, ninety, nodded, nought, oh, okay, on, ooh, out, pence, percent, political, presence, provides, put, really, reckon, say, seemed, seriously, sixty, smiled, so, social, somebody, system, take, talking, than, the, they, thing, think, thirteen, though, thus, time, tt, tv, twenty, uk, under, urgh, us, usa, wants, was, we, who, with, world, yeah, yes, you, your

The data are provided in the form of a term-document matrix, where rows denote the 150 items and columns denote the 500 texts in the corpus. Four items do not occur in ICE-GB (*aye, corp, ltd, tt*). These are included in the term-document matrix with frequencies of 0 for all texts.

The first row of the term-document matrix gives the length of the text (i.e. number of word tokens).

Source

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Nelson, Gerald, Sean Wallis and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

`biber150_ice_gb_genre` *Distribution of Biber et al.'s (2016) 150 lexical items in ICE-GB (term-document matrix), using genres as corpus parts*

Description

This dataset contains text-level frequencies for ICE-GB (Nelson et al. 2002) for a set of 150 word forms. The list of items was compiled by Biber et al. (2016) for methodological purposes, that is, to study the behavior of dispersion measures in different distributional settings. The items are intended to cover a broad range of frequency and dispersion levels.

Usage

```
biber150_ice_gb_genre
```

Format

`biber150_ice_gb_genre`**:**

A matrix with 151 rows and 32 columns

rows Size of the genre (`word_count`), followed by set of 150 items in alphabetical order (*a, able, ..., you, your*)

columns 32 genres, ordered alphabetically ("acad_humanities", "acad_natural_sciences", "acad_social_sciences", ... ,"student_essays", "unscripted_speeches"))

Details

While Biber et al. (2016: 446) used 153 target items, the 150 word forms included in the present data set correspond to the slightly narrower selection of forms used in Burch et al. (2017: 214-216). These 150 word forms are listed next, in alphabetical order:

a, able, actually, after, against, ah, aha, all, among, an, and, another, anybody, at, aye, be, became, been, began, bet, between, bloke, both, bringing, brought, but, charles, claimed, cor, corp, cos, da, day, decided, did, do, doo, during, each, economic, eh, eighty, england, er, erm, etcetera, everybody, fall, fig, for, forty, found, from, full, get, government, ha, had, has, have, having, held, hello, himself, hm, however, hundred, i, ibm, if, important, in, inc, including, international, into, it, just, know, large, later, latter, let, life, ltd, made, may, methods, mhm, minus, mm, most, mr, mum, new, nineteen, ninety, nodded, nought, oh, okay, on, ooh, out, pence, percent, political, presence, provides, put, really, reckon, say, seemed, seriously, sixty, smiled, so, social, somebody, system, take, talking, than, the, they, thing, think, thirteen, though, thus, time, tt, tv, twenty, uk, under, urgh, us, usa, wants, was, we, who, with, world, yeah, yes, you, your

The data are provided in the form of a term-document matrix, where rows denote the 150 items and columns denote the 32 genres in the corpus. Four items do not occur in ICE-GB (*aye, corp, ltd, tt*). These are included in the term-document matrix with frequencies of 0 for all texts.

The first row of the term-document matrix gives the size of the genre (i.e. number of word tokens).

Source

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Nelson, Gerald, Sean Wallis and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

`biber150_ice_gb_macro_genre`

Distribution of Biber et al.'s (2016) 150 lexical items in ICE-GB (term-document matrix), using macro-genres as corpus parts

---

Description

This dataset contains text-level frequencies for ICE-GB (Nelson et al. 2002) for a set of 150 word forms. The list of items was compiled by Biber et al. (2016) for methodological purposes, that is, to study the behavior of dispersion measures in different distributional settings. The items are intended to cover a broad range of frequency and dispersion levels.

Usage

```
biber150_ice_gb_macro_genre
```

Format

`biber150_ice_gb_macro_genre`**:**

A matrix with 151 rows and 12 columns

rows Size of the macro-genre (`word_count`), followed by set of 150 items in alphabetical order (*a, able, ..., you, your*)

columns 12 macro-genres, ordered alphabetically ("academic_writing", "creative_writing", "instructional_writing", ... ,"student_writing", "unscripted_monologues"))

Details

While Biber et al. (2016: 446) used 153 target items, the 150 word forms included in the present data set correspond to the slightly narrower selection of forms used in Burch et al. (2017: 214-216). These 150 word forms are listed next, in alphabetical order:

a, able, actually, after, against, ah, aha, all, among, an, and, another, anybody, at, aye, be, became, been, began, bet, between, bloke, both, bringing, brought, but, charles, claimed, cor, corp, cos, da, day, decided, did, do, doo, during, each, economic, eh, eighty, england, er, erm, etcetera, everybody, fall, fig, for, forty, found, from, full, get, government, ha, had, has, have, having, held, hello, himself, hm, however, hundred, i, ibm, if, important, in, inc, including, international, into, it, just, know, large, later, latter, let, life, ltd, made, may, methods, mhm, minus, mm, most, mr, mum, new, nineteen, ninety, nodded, nought, oh, okay, on, ooh, out, pence, percent, political, presence, provides, put, really, reckon, say, seemed, seriously, sixty, smiled, so, social, somebody, system, take, talking, than, the, they, thing, think, thirteen, though, thus, time, tt, tv, twenty, uk, under, urgh, us, usa, wants, was, we, who, with, world, yeah, yes, you, your

The data are provided in the form of a term-document matrix, where rows denote the 150 items and columns denote the 12 macro-genres in the corpus. Four items do not occur in ICE-GB (*aye, corp, ltd, tt*). These are included in the term-document matrix with frequencies of 0 for all texts.

The first row of the term-document matrix gives the size of the genre (i.e. number of word tokens).

Source

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Nelson, Gerald, Sean Wallis and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

`biber150_spokenBNC1994` *Distribution of Biber et al.'s (2016) 150 lexical items in the Spoken BNC1994 (term-document matrix)*

Description

This dataset contains speaker-level frequencies for the demographically sampled part of the Spoken BNC1994 (Crowdy 1995) for a set of 150 word forms. The list of items was compiled by Biber et al. (2016) for methodological purposes, that is, to study the behavior of dispersion measures in different distributional settings. The items are intended to cover a broad range of frequency and dispersion levels.

Usage

```
biber150_spokenBNC1994
```

Format

`biber150_spokenBNC1994`**:**

A matrix with 151 rows and 1,017 columns

rows Total number of words by speaker (`word_count`), followed by set of 150 items in alphabetical order (*a, able, ..., you, your*)

columns 1,405 speakers, ordered by ID ("PS002","PS003", ... , "PS6SM", "PS6SN"))

Details

While Biber et al. (2016: 446) used 153 target items, the 150 word forms included in the present data set correspond to the slightly narrower selection of forms used in Burch et al. (2017: 214-216). These 150 word forms are listed next, in alphabetical order:

a, able, actually, after, against, ah, aha, all, among, an, and, another, anybody, at, aye, be, became, been, began, bet, between, bloke, both, bringing, brought, but, charles, claimed, cor, corp, cos, da, day, decided, did, do, doo, during, each, economic, eh, eighty, england, er, erm, etcetera, everybody, fall, fig, for, forty, found, from, full, get, government, ha, had, has, have, having, held, hello, himself, hm, however, hundred, i, ibm, if, important, in, inc, including, international, into, it, just, know, large, later, latter, let, life, ltd, made, may, methods, mhm, minus, mm, most, mr, mum, new, nineteen, ninety, nodded, nought, oh, okay, on, ooh, out, pence, percent, political, presence, provides, put, really, reckon, say, seemed, seriously, sixty, smiled, so, social, somebody, system, take, talking, than, the, they, thing, think, thirteen, though, thus, time, tt, tv, twenty, uk, under, urgh, us, usa, wants, was, we, who, with, world, yeah, yes, you, your

The data are provided in the form of a term-document matrix, where rows denote the 150 items and columns denote 1,017 speakers in the demographically sampled part of the corpus. This dataset only includes speakers for whom information on both age and sex are available.

The first row of the term-document matrix gives the total number of words (i.e. number of word tokens) the speaker contributed to the corpus.

Source

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Crowdy, Steve. 1995. The BNC spoken corpus. In Geoffrey Leech, Greg Myers & Jenny Thomas (eds.), *Spoken English on Computer: Transcription, Mark-Up and Annotation*, 224–234. Harlow: Longman.

`biber150_spokenBNC2014`

Distribution of Biber et al.'s (2016) 150 lexical items in the Spoken BNC2014 (term-document matrix)

Description

This dataset contains speaker-level frequencies for the Spoken BNC2014 (Love et al. 2017) for a set of 150 word forms. The list of items was compiled by Biber et al. (2016) for methodological purposes, that is, to study the behavior of dispersion measures in different distributional settings. The items are intended to cover a broad range of frequency and dispersion levels.

Usage

```
biber150_spokenBNC2014
```

Format

`biber150_spokenBNC2014`**:**

A matrix with 151 rows and 668 columns

rows Total number of words by speaker (`word_count`), followed by set of 150 items in alphabetical order (*a, able, ..., you, your*)

columns 668 speakers, ordered by ID ("S0001","S0002", ... , "S0691", "S0692"))

Details

While Biber et al. (2016: 446) used 153 target items, the 150 word forms included in the present data set correspond to the slightly narrower selection of forms used in Burch et al. (2017: 214-216). These 150 word forms are listed next, in alphabetical order:

a, able, actually, after, against, ah, aha, all, among, an, and, another, anybody, at, aye, be, became, been, began, bet, between, bloke, both, bringing, brought, but, charles, claimed, cor, corp, cos, da, day, decided, did, do, doo, during, each, economic, eh, eighty, england, er, erm, etcetera, everybody, fall, fig, for, forty, found, from, full, get, government, ha, had, has, have, having, held, hello, himself, hm, however, hundred, i, ibm, if, important, in, inc, including, international, into, it,

just, know, large, later, latter, let, life, ltd, made, may, methods, mhm, minus, mm, most, mr, mum, new, nineteen, ninety, nodded, nought, oh, okay, on, ooh, out, pence, percent, political, presence, provides, put, really, reckon, say, seemed, seriously, sixty, smiled, so, social, somebody, system, take, talking, than, the, they, thing, think, thirteen, though, thus, time, tt, tv, twenty, uk, under, urgh, us, usa, wants, was, we, who, with, world, yeah, yes, you, your

The data are provided in the form of a term-document matrix, where rows denote the 150 items and columns denote the 668 speakers in the corpus. Speakers with the label "UNKFEMALE", "UNKMALE", and "UNKMULTI" are not included in the dataset.

The first row of the term-document matrix gives the total number of words (i.e. number of word tokens) the speaker contributed to the corpus.

Source

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319–344.

---

disp *Calculate parts-based dispersion measures*

Description

This function calculates a number of parts-based dispersion measures and allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp(
  subfreq,
  partsize,
  directionality = "conventional",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

subfreq	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
partsize	A numeric vector specifying the size of the corpus parts
directionality	Character string indicating the directionality of scaling. See details below. Possible values are "conventional" (default) and "gries"
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE
digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)
verbose	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE
print_score	Logical. Whether the dispersion score should be printed to the console; default is TRUE
suppress_warning	Logical. Whether warning messages should be suppressed; default is FALSE

Details

This function calculates dispersion measures based on two vectors: a set of subfrequencies (number of occurrences of the item in each corpus part) and a matching set of part sizes (the size of the corpus parts, i.e. number of word tokens).

- Directionality: The scores for all measures range from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest

value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.

- To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).
- To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()` and `vignette("frequency-adjustment")`.

The following measures are computed, listed in chronological order (see details below):

- R_{rel} (Keniston 1920)
- D (Juilland & Chang-Rodriguez 1964)
- D_2 (Carroll 1970)
- S (Rosengren 1971)
- D_P (Gries 2008; modification: Egbert et al. 2020)
- D_A (Burch et al. 2017)
- D_{KL} (Gries 2024)

In the formulas given below, the following notation is used:

- k the number of corpus parts
- T_i the absolute subfrequency in part i
- t_i a proportional quantity; the subfrequency in part i divided by the total number of occurrences of the item in the corpus (i.e. the sum of all subfrequencies)
- W_i the absolute size of corpus part i
- w_i a proportional quantity; the size of corpus part i divided by the size of the corpus (i.e. the sum of the part sizes)
- R_i the normalized subfrequency in part i, i.e. the subfrequency divided by the size of the corpus part
- r_i a proportional quantity; the normalized subfrequency in part i divided by the sum of all normalized subfrequencies
- N corpus frequency, i.e. the total number of occurrence of the item in the corpus

Note that the formulas cited below differ in their scaling, i.e. whether 1 reflects an even or an uneven distribution. In the current function, this behavior is overridden by the argument `directionality`. The specific scaling used in the formulas below is therefore irrelevant.

R_{rel} refers to the relative range, i.e. the proportion of corpus parts containing at least one occurrence of the item.

D denotes Juilland's D and is calculated as follows (this formula uses conventional scaling); $\bar{R}_i$ refers to the average over the normalized subfrequencies:

$1 - \sqrt{\frac{\sum_{i=1}^{k}(R_i - \bar{R}_i)^2}{k}} \times \frac{1}{\bar{R}_i\sqrt{k-1}}$

D_2 denotes the index proposed by Carroll (1970); the following formula uses conventional scaling:

$\frac{\sum_i^k r_i \log_2 \frac{1}{r_i}}{\log_2 k}$

S is the dispersion measure proposed by Rosengren (1971); the formula uses conventional scaling:

$\frac{(\sum_i^k r_i\sqrt{w_i T_i}}{N}$

D_P represents Gries's deviation of proportions; the following formula is the modified version suggested by Egbert et al. (2020: 99); it implements conventional scaling (0 = uneven, 1 = even) and the notation $min\{w_i : t_i > 0\}$ refers to the w_i value among those corpus parts that include at least one occurrence of the item.

$1 - \frac{\sum_i^k |t_i - w_i|}{2} \times \frac{1}{1 - min\{w_i : t_i > 0\}}$

D_A is a measure introduced into dispersion analysis by Burch et al. (2017). The following formula is the one used by Egbert et al. (2020: 98); it relies on normalized frequencies and therefore works with corpus parts of different size. The formula represents conventional scaling (0 = uneven, 1 = even):

$1 - \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}|R_i - R_j|}{\frac{k(k-1)}{2}} \times \frac{1}{2\frac{\sum_i^k R_i}{k}}$

The current function uses a formula that may be found in Wilcox (1973: 343). It relies on the proportional r_i values instead of the normalized subfrequencies R_i:

$1 - \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}|r_i - r_j|}{k-1}$

Since this formula is computationally expensive, the function actually uses the computational shortcut given in Wilcox (1973: 343). Critically, the proportional quantities r_i must first be sorted in decreasing order. Only after this rearrangement can the shortcut version be applied. We will refer to this rearranged version of r_i as r_i^{sorted}:

$\frac{2\left(\sum_{i=1}^{k}(i \times r_i^{sorted}) - 1\right)}{k-1}$ (Wilcox 1973: 343)

D_{KL} refers to a measure proposed by Gries (2020, 2021); for standardization, it uses the odds-to-probability transformation (Gries 2024: 90) and represents Gries scaling (0 = even, 1 = uneven):

$\frac{\sum_i^k t_i \log_2 \frac{t_i}{w_i}}{1 + \sum_i^k t_i \log_2 \frac{t_i}{w_i}}$

Value

A numeric vector of seven dispersion scores

Author(s)

Lukas Soenning

References

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216. doi:10.1558/jrds.33066

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115. doi:10.1075/ijcl.18010.egb

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 99–118. New York: Springer. doi:10.1007/9783030-462161_5

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2). 1–33. doi:10.32714/ricl.09.02.02

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Keniston, Hayward. 1920. Common words in Spanish. *Hispania* 3(2). 85–96. doi:10.2307/331305

Lijffijt, Jefrey & Stefan Th. Gries. 2012. Correction to Stefan Th. Gries' 'Dispersions and adjusted frequencies in corpora'. *International Journal of Corpus Linguistics* 17(1). 147–149. doi:10.1075/ijcl.17.1.08lij

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

See Also

For finer control over the calculation of several dispersion measures:

- `disp_R()` for $Range$
- `disp_DP()` for D_P
- `disp_DA()` for D_A
- `disp_DKL()` for D_{KL}

Examples

```
disp_DP(
  subfreq = c(0,0,1,2,5),
  partsize = rep(1000, 5),
  directionality = "conventional",
  freq_adjust = FALSE)
```

disp_DA *Calculate the dispersion measure D_A*

Description

This function calculates the dispersion measure D_A. It allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also provides the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_DA(
  subfreq,
  partsize,
  directionality = "conventional",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

subfreq	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
partsize	A numeric vector specifying the size of the corpus parts
directionality	Character string indicating the directionality of scaling. See details below. Possible values are "conventional" (default) and "gries"
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE
digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)
verbose	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE
print_score	Logical. Whether the dispersion score should be printed to the console; default is TRUE
suppress_warning	Logical. Whether warning messages should be suppressed; default is FALSE

Details

The function calculates the dispersion measure D_A based on a set of subfrequencies (number of occurrences of the item in each corpus part) and a matching set of part sizes (the size of the corpus parts, i.e. number of word tokens). The function uses the shortcut formula ("computational" procedure) given in Wilcox (1973: 343), where D_A is referred to as MDA.

- Directionality: D_A ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.
  - To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).
  - To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()` and `vignette("frequency-adjustment")`.

In the formulas given below, the following notation is used:

- k the number of corpus parts
- R_i the normalized subfrequency in part i, i.e. the number of occurrences of the item divided by the size of the part
- r_i a proportional quantity; the normalized subfrequency in part i (R_i) divided by the sum of all normalized subfrequencies

The basic formula for D_A (see Wilcox 1973: 329, 343; Burch et al. 2017: 194; Egbert et al. 2020: 98) can be applied to absolute frequencies or normalized frequencies. For dispersion analysis, absolute frequencies only make sense if the corpus parts are identical in size. Wilcox (1973: 343, 'MDA', column 1 and 2) gives both variants of the basic version. The first use of D_A for corpus-linguistic dispersion analysis appears in Burch et al. (2017: 194), a paper that deals with equal-sized parts and therefore uses the variant for absolute frequencies. Egbert et al. (2020: 98) rely

on the variant using normalized frequencies. Since this variant of the basic version of D_A works irrespective of the length of the corpus parts (equal or variable), we will only give this version of the formula. Note that while the formula represents conventional scaling (0 = uneven, 1 = even), in the current function the directionality is controlled separately using the argument `directionality`.

$1 - \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |R_i - R_j|}{\frac{k(k-1)}{2}} \times \frac{1}{2\frac{\sum_{i}^{k} R_i}{k}}$ (Egbert et al. 2020: 98)

The function uses a different version of the same formula, which relies on the proportional r_i values instead of the normalized subfrequencies R_i. This version yields identical results; the r_i quantities are also the key to using the computational shortcut given in Wilcox (1973: 343), on which the calculations in the `{tlda}` package rely. This is the basic formula for D_A using r_i instead of R_i values:

$1 - \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |r_i - r_j|}{k-1}$ (Wilcox 1973: 343; see also Soenning 2022)

Functions for D_A in the `{tlda}` package use the computational shortcut given in Wilcox (1973: 343). Critically, the proportional quantities r_i must first be sorted in decreasing order. Only after this rearrangement can the shortcut version be applied. We will refer to this rearranged version of r_i as r_i^{sorted}:

$\frac{2\left(\sum_{i=1}^{k} (i \times r_i^{sorted}) - 1\right)}{k-1}$ (Wilcox 1973: 343)

Value

A numeric value

Author(s)

Lukas Soenning

References

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216. doi:10.1558/jrds.33066

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115. doi:10.1075/ijcl.18010.egb

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Soenning, Lukas. 2022. Evaluation of text-level measures of lexical dispersion: Robustness and consistency. *PsyArXiv preprint*. https://osf.io/preprints/psyarxiv/h9mvs/

Wilcox, Allen R. 1973. Indices of qualitative variation and political measurement. *The Western Political Quarterly* 26 (2). 325–343. doi:10.2307/446831

Examples

```
disp_DA(
  subfreq = c(0,0,1,2,5),
  partsize = rep(1000, 5),
  directionality = "conventional",
  freq_adjust = FALSE)
```

---

disp_DA_tdm *Calculate the dispersion measure D_A for a term-document matrix*

Description

This function calculates the dispersion measure D_A. It allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also provides the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_DA_tdm(
  tdm,
  row_partsize = "first",
  directionality = "conventional",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  add_frequency = TRUE,
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_scores = TRUE
)
```

Arguments

tdm	A term-document matrix, where rows represent items and columns represent corpus parts; must also contain a row giving the size of the corpus parts (first or last row in the term-document matrix)
row_partsize	Character string indicating which row in the term-document matrix contains the size of the corpus parts. Possible values are "first" (default) and "last"

`directionality` Character string indicating the directionality of scaling. See details below. Possible values are `"conventional"` (default) and `"gries"`

`freq_adjust` Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is `FALSE`

`freq_adjust_method`
Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are `"even"` (default) and `"pervasive"`

`add_frequency` Logical. Whether to add a column that gives the total number of occurrences of the item across a corpus parts; default is `TRUE`

`unit_interval` Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is `TRUE`

`digits` Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)

`verbose` Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is `TRUE`

`print_scores` Logical. Whether the dispersion scores should be printed to the console; default is `TRUE`

Details

This function takes as input a term-document matrix and returns, for each item (i.e. each row) the dispersion measure D_A. The rows in the input matrix represent the items, and the columns the corpus parts. Importantly, the term-document matrix must include an additional row that records the size of the corpus parts. For a proper term-document matrix, which includes all items that appear in the corpus, this can be added as a column margin, which sums the frequencies in each column. If the matrix only includes a selection of items drawn from the corpus, this information cannot be derived from the matrix and must be provided as a separate row. The function uses the shortcut formula ("computational" procedure) given in Wilcox (1973: 343), where D_A is referred to as MDA.

- Directionality: D_A ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = 'gries'` to choose this option.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022, 2024). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different

way, based on the distributional features pervasiveness (`pervasive`) or evenness (`even`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.

- To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`pervasive`), or they are assigned to the smallest corpus part(s) (`even`).
- To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`pervasive`), or they are allocated to corpus parts in proportion to their size (`even`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()`.

In the formulas given below, the following notation is used:

- k the number of corpus parts
- R_i the normalized subfrequency in part i, i.e. the number of occurrences of the item divided by the size of the part
- r_i a proportional quantity; the normalized subfrequency in part i (R_i) divided by the sum of all normalized subfrequencies

The basic formula for D_A (see Wilcox 1973: 329, 343; Burch et al. 2017: 194; Egbert et al. 2020: 98) can be applied to absolute frequencies or normalized frequencies. For dispersion analysis, absolute frequencies only make sense if the corpus parts are identical in size. Wilcox (1973: 343, 'MDA', column 1 and 2) gives both variants of the basic version. The first use of D_A for corpus-linguistic dispersion analysis appears in Burch et al. (2017: 194), a paper that deals with equal-sized parts and therefore uses the variant for absolute frequencies. Egbert et al. (2020: 98) rely on the variant using normalized frequencies. Since this variant of the basic version of D_A works irrespective of the length of the corpus parts (equal or variable), we will only give this version of the formula. Note that while the formula represents conventional scaling (0 = uneven, 1 = even), in the current function the directionality is controlled separately using the argument `directionality`.

$1 - \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |R_i - R_j|}{\frac{k(k-1)}{2}} \times \frac{1}{2\frac{\sum_{i}^{k} R_i}{k}}$ (Egbert et al. 2020: 98)

The function uses a different version of the same formula, which relies on the proportional r_i values instead of the normalized subfrequencies R_i. This version yields identical results; the r_i quantities are also the key to using the computational shortcut given in Wilcox (1973: 343), on which the calculations in the `{tlda}` package rely. This is the basic formula for D_A using r_i instead of R_i values:

$1 - \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |r_i - r_j|}{k-1}$ (Wilcox 1973: 343; see also Soenning 2022)

Functions for D_A in the `{tlda}` package use the computational shortcut given in Wilcox (1973: 343). Critically, the proportional quantities r_i must first be sorted in decreasing order. Only after this rearrangement can the shortcut version be applied. We will refer to this rearranged version of r_i as r_i^{sorted}:

$\frac{2\left(\sum_{i=1}^{k}(i \times r_i^{sorted}) - 1\right)}{k-1}$ (Wilcox 1973: 343)

Value

A data frame with one row per item

Author(s)

Lukas Soenning

References

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216. doi:10.1558/jrds.33066

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115. doi:10.1075/ijcl.18010.egb

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Soenning, Lukas. 2022. Evaluation of text-level measures of lexical dispersion: Robustness and consistency. *PsyArXiv preprint*. https://osf.io/preprints/psyarxiv/h9mvs/

Wilcox, Allen R. 1973. Indices of qualitative variation and political measurement. *The Western Political Quarterly* 26 (2). 325–343. doi:10.2307/446831

Examples

```
disp_DA_tdm(
  tdm = biber150_spokenBNC2014[1:20,],
  row_partsize = "first",
  directionality = "conventional",
  freq_adjust = FALSE)
```

`disp_DKL` *Calculate the dispersion measure D_KL*

Description

This function calculates the dispersion measure D_{KL}, which is based on the Kullback-Leibler divergence (Gries 2020, 2021, 2024). It offers three options for standardization to the unit interval [0,1] (see Gries 2024: 90-92) and allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_DKL(
  subfreq,
  partsize,
  directionality = "conventional",
  standardization = "o2p",
  custom_base = NULL,
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

`subfreq`	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
`partsize`	A numeric vector specifying the size of the corpus parts
`directionality`	Character string indicating the directionality of scaling. See details below. Possible values are `"conventional"` (default) and `"gries"`
`standardization`	Character string indicating which standardization method to use. See details below. Possible values are `"o2p"` (default), `"base_e"`, `"base_2"`, and `"custom"`.
`custom_base`	A numeric value specifying the custom base for standardization; only work with `standardization = "custom"`; see details below
`freq_adjust`	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is `FALSE`
`freq_adjust_method`	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are `"even"` (default) and `"pervasive"`
`unit_interval`	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is `TRUE`
`digits`	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)

`verbose` Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is `TRUE`

`print_score` Logical. Whether the dispersion score should be printed to the console; default is `TRUE`

`suppress_warning` Logical. Whether warning messages should be suppressed; default is `FALSE`

Details

The function calculates the dispersion measure D_{KL} based on a set of subfrequencies (number of occurrences of the item in each corpus part) and a matching set of part sizes (the size of the corpus parts, i.e. number of word tokens).

- Directionality: D_{KL} ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option.
- Standardization: Irrespective of the directionality of scaling, three ways of standardizing the Kullback-Leibler divergence to the unit interval [0;1] are mentioned in Gries (2024: 90-92). The choice between these transformations can have an appreciable effect on the standardized dispersion score. In Gries (2020: 103-104), the Kullback-Leibler divergence is not standardized. In Gries (2021: 20), the transformation `"base_e"` is used (see (1) below), and in Gries (2024), the default strategy is `"o2p"`, the odds-to-probability transformation (see (3) below).
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.
  - To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).
  - To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()` and `vignette("frequency-adjustment")`.

In the formulas given below, the following notation is used:

- t_i a proportional quantity; the subfrequency in part i divided by the total number of occurrences of the item in the corpus (i.e. the sum of all subfrequencies)
- w_i a proportional quantity; the size of corpus part i divided by the size of the corpus (i.e. the sum of the part sizes)

The first step is to calculate the Kullback-Leibler divergence based on the proportional subfrequencies (t_i) and the size of the corpus parts (w_i):

$KLD = \sum_i^k t_i \log_2 \frac{t_i}{w_i}$ with $\log_2(0) = 0$

This KLD score is then standardized (i.e. transformed) to the conventional unit interval [0,1]. Three options are discussed in Gries (2024: 90-92). The following formulas represent Gries scaling (0 = even, 1 = uneven):

(1) e^{-KLD} (Gries 2021: 20), represented by the value `"base_e"`

(2) 2^{-KLD} (Gries 2024: 90), represented by the value" `"base_2"`

(3) $\frac{KLD}{1+KLD}$ (Gries 2024: 90), represented by the value `"o2p"` (default)

A fourth option is available which allows the user to select a custom base for standardization (i.e. a value other than e (`"base_e"`) and 2 (`"base_2"`)). If the argument `standardization` is set to `"custom"`, a numeric value must be supplied to the argument `custom_base`.

(4) b^{-KLD} (with b representing a numeric base) represented by the value" `"custom"` and `custom_base = b`

Value

A numeric value

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
disp_DKL(
  subfreq = c(0,0,1,2,5),
  partsize = rep(1000, 5),
  standardization = "base_e",
  directionality = "conventional")
```

`disp_DKL_tdm` *Calculate the dispersion measure D_KL for a term-document matrix*

Description

This function calculates the dispersion measure D_{KL}, which is based on the Kullback-Leibler divergence (Gries 2020, 2021, 2024). It offers three different options for standardization to the unit interval [0,1] (see Gries 2024: 90-92) and allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_DKL_tdm(
  tdm,
  row_partsize = "first",
  directionality = "conventional",
  standardization = "o2p",
  custom_base = NULL,
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  add_frequency = TRUE,
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_scores = TRUE
)
```

Arguments

`tdm`	A term-document matrix, where rows represent items and columns represent corpus parts; must also contain a row giving the size of the corpus parts (first or last row in the term-document matrix)
`row_partsize`	Character string indicating which row in the term-document matrix contains the size of the corpus parts. Possible values are `"first"` (default) and `"last"`
`directionality`	Character string indicating the directionality of scaling. See details below. Possible values are `"conventional"` (default) and `"gries"`

`standardization`
: Character string indicating which standardization method to use. See details below. Possible values are `"o2p"` (default), `"base_e"`, `"base_2"`, and `"custom"`.

`custom_base`
: A numeric value specifying the custom base for standardization; only work with `standardization = "custom"`; see details below

`freq_adjust`
: Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is `FALSE`

`freq_adjust_method`
: Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are `"even"` (default) and `"pervasive"`

`add_frequency`
: Logical. Whether to add a column that gives the total number of occurrences of the item across a corpus parts; default is `TRUE`

`unit_interval`
: Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is `TRUE`

`digits`
: Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)

`verbose`
: Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is `TRUE`

`print_scores`
: Logical. Whether the dispersion scores should be printed to the console; default is `TRUE`

Details

This function takes as input a term-document matrix and returns, for each item (i.e. each row) the dispersion measure D_{KL}. The rows in the input matrix represent the items, and the columns the corpus parts. Importantly, the term-document matrix must include an additional row that records the size of the corpus parts. For a proper term-document matrix, which includes all items that appear in the corpus, this can be added as a column margin, which sums the frequencies in each column. If the matrix only includes a selection of items drawn from the corpus, this information cannot be derived from the matrix and must be provided as a separate row.

- Directionality: D_{KL} ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = 'gries'` to choose this option.
- Standardization: Irrespective of the directionality of scaling, three ways of standardizing the Kullback-Leibler divergence to the unit interval [0;1] are mentioned in Gries (2024: 90-92). The choice between these transformations can have an appreciable effect on the standardized dispersion score. In Gries (2020: 103-104), the Kullback-Leibler divergence is not standardized. In Gries (2021: 20), the transformation `'base_e'` is used (see (1) below), and in Gries (2024), the default strategy is `'o2p'`, the odds-to-probability transformation (see (3) below).
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is

set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`pervasive`) or evenness (`even`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.

- To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`pervasive`), or they are assigned to the smallest corpus part(s) (`even`).
- To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`pervasive`), or they are allocated to corpus parts in proportion to their size (`even`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()`.

In the formulas given below, the following notation is used:

- t_i a proportional quantity; the subfrequency in part i divided by the total number of occurrences of the item in the corpus (i.e. the sum of all subfrequencies)
- w_i a proportional quantity; the size of corpus part i divided by the size of the corpus (i.e. the sum of the part sizes)

The first step is to calculate the Kullback-Leibler divergence based on the proportional subfrequencies (t_i) and the size of the corpus parts (w_i):

$KLD = \sum_i^k t_i \log_2 \frac{t_i}{w_i}$ with $\log_2(0) = 0$

This KLD score is then standardized (i.e. transformed) to the conventional unit interval [0,1]. Three options are discussed in Gries (2024: 90-92). The following formulas represents Gries scaling (0 = even, 1 = uneven):

(1) e^{-KLD} (Gries 2021: 20), represented by the value `'base_e'`

(2) 2^{-KLD} (Gries 2024: 90), represented by the value `'base_2'`

(3) $\frac{KLD}{1+KLD}$ (Gries 2024: 90), represented by the value `'o2p'` (default)

Value

A data frame with one row per item

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
disp_DKL_tdm(
  tdm = biber150_spokenBNC2014[1:20,],
  row_partsize = "first",
  standardization = "base_e",
  directionality = "conventional")
```

`disp_DMB` *Calculate the dispersion measure* D~MB~

Description

This function calculates *D~MB~*, a generalized version of the Poisson-based dispersion measure *MB* proposed by Nelson (2025). It allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution, and it returns confidence intervals for *D~MB~*.

Usage

```
disp_DMB(
  subfreq,
  partsize,
  directionality = "conventional",
  conf_int = FALSE,
  conf_level = 0.95,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

`subfreq`	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
`partsize`	A numeric vector specifying the size of the corpus parts
`directionality`	Character string indicating the directionality of scaling. See details below. Possible values are `"conventional"` (default) and `"gries"`
`conf_int`	Logical. Whether a (profile likelihood) confidence interval should be computed; default: `FALSE`
`conf_level`	Scalar giving the confidence level; default `0.95` for a 95% CI
`digits`	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)
`verbose`	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is `TRUE`
`print_score`	Logical. Whether the dispersion score should be printed to the console; default is `TRUE`
`suppress_warning`	Logical. Whether warning messages should be suppressed; default is `FALSE`

Details

The function calculates the dispersion measure *D~MB~* based on a set of subfrequencies (number of occurrences of the item in each corpus part) and a matching set of part sizes (the size of the corpus parts, i.e. number of word tokens). *D~MB~* can be considered a generalization of the method proposed by Nelson (2025). In contrast to the original measure, *MB*, *D~MB~* works with a pre-determined set of corpus parts, which may also differ in size. To provide this additional flexibility, *D~MB~* is constructed based on a Poisson regression model that considers the corpus parts as observations and allows them to differ in length through its incorporation of an offset parameter.

- Directionality: *D~MB~* ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option.

Value

A vector of numeric values

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Nelson, Robset N. Jr. 2025. Groundhog Day is not a good model for corpus dispersion. Journal of Quantitative Linguistics 32(2). 103–127. doi:10.1080/09296174.2024.2423415

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
disp_DMB(
  subfreq = c(0,0,1,2,5),
  partsize = rep(1000, 5),
  directionality = "conventional",
  conf_int = TRUE)
```

---

disp_DP *Calculate Gries's* deviation of proportions

Description

This function calculates Gries's dispersion measure DP (deviation of proportions). It offers three different formulas and allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_DP(
  subfreq,
  partsize,
  directionality = "conventional",
  formula = "egbert_etal_2020",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

subfreq	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
partsize	A numeric vector specifying the size of the corpus parts
directionality	Character string indicating the directionality of scaling. See details below. Possible values are "conventional" (default) and "gries"
formula	Character string indicating which formula to use for the calculation of DP. See details below. Possible values are "egbert_etal_2020" (default), "gries_2008", "lijffit_gries_2012".
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE
digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)
verbose	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE
print_score	Logical. Whether the dispersion score should be printed to the console; default is TRUE
suppress_warning	Logical. Whether warning messages should be suppressed; default is FALSE

Details

The function calculates the dispersion measure DP based on a set of subfrequencies (number of occurrences of the item in each corpus part) and a matching set of part sizes (the size of the corpus parts, i.e. number of word tokens).

- Directionality: DP ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option.
- Formula: Irrespective of the directionality of scaling, four formulas for DP exist in the literature (see below for details). This is because the original version proposed by Gries (2008: 415), which is commonly denoted as DP (and here referenced by the value "gries_2008") does not always reach its theoretical limits of 0 and 1. For this reason, modifications have been suggested, starting with Gries (2008: 419) himself, who referred to this version as $DPnorm$. This version is not implemented in the current package, because Lijffit & Gries (2012) updated $DPnorm$ to ensure that it also works as intended when corpus parts differ in size; this version is represented by the value "lijffit_gries_2012" and often denoted using subscript notation DP_{norm} . Finally, Egbert et al. (2020: 99) suggest a further modification to ensure proper behavior in settings where the item occurs in only one corpus part. They label this version D_P. In the current function, it is the default and represented by the value "egbert_etal_2020".

- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.
  - To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).
  - To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()`.

In the formulas given below, the following notation is used:

- k the number of corpus parts
- t_i a proportional quantity; the subfrequency in part i divided by the total number of occurrences of the item in the corpus (i.e. the sum of all subfrequencies)
- w_i a proportional quantity; the size of corpus part i divided by the size of the corpus (i.e. the sum of thea part sizes)

The value `"gries_2008"` implements the original version proposed by Gries (2008: 415). Note that while the following formula represents Gries scaling (0 = even, 1 = uneven), in the current function the directionality is controlled separately using the argument `directionality`.

$\frac{\sum_i^k |t_i - w_i|}{2}$ (Gries 2008)

The value `"lijffit_gries_2012"` implements the modified version described by Lijffit & Gries (2012). Again, the following formula represents Gries scaling (0 = even, 1 = uneven), but the directionality is handled separately in the current function. The notation $min\{w_i\}$ refers to the w_i value of the smallest corpus part.

$\frac{\sum_i^k |t_i - w_i|}{2} \times \frac{1}{1 - min\{w_i\}}$ (Lijffijt & Gries 2012)

The value `"egbert_etal_2020"` (default) selects the modification suggested by Egbert et al. (2020: 99). The following formula represents conventional scaling (0 = uneven, 1 = even). The notation $min\{w_i : t_i > 0\}$ refers to the w_i value among those corpus parts that include at least one occurrence of the item.

$1 - \frac{\sum_i^k |t_i - w_i|}{2} \times \frac{1}{1 - min\{w_i : t_i > 0\}}$ (Egbert et al. 2020)

Value

A numeric value

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115. doi:10.1075/ijcl.18010.egb

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Lijffijt, Jefrey & Stefan Th. Gries. 2012. Correction to Stefan Th. Gries' 'Dispersions and adjusted frequencies in corpora'. *International Journal of Corpus Linguistics* 17(1). 147–149. doi:10.1075/ijcl.17.1.08lij

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
disp_DP(
  subfreq = c(0,0,1,2,5),
  partsize = rep(1000, 5),
  directionality = "conventional",
  formula = "gries_2008",
  freq_adjust = FALSE)
```

---

disp_DP_boot *Bootstrap (and weight) Gries's* deviation of proportions

---

Description

This function offers facilities for bootstrapping and weighting Gries's dispersion measure DP (deviation of proportions). In addition to the full functionality offered by the function `disp_DP()`, it can be used to obtain weighted dispersion scores as well as bootstrap confidence intervals.

Usage

```
disp_DP_boot(
  subfreq,
  partsize,
  n_boot = 500,
  boot_ci = FALSE,
  conf_level = 0.95,
  return_distribution = FALSE,
  partweight = NULL,
  directionality = "conventional",
  formula = "egbert_etal_2020",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

subfreq	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
partsize	A numeric vector specifying the size of the corpus parts
n_boot	Integer value specifying the number of bootstrap samples to draw; default: 500
boot_ci	Logical. Whether a percentile bootstrap confidence interval should be computed; default: FALSE
conf_level	Scalar giving the confidence level; default 0.95 for a 95% percentile CI
return_distribution	Logical. Whether the function should return a vector of all n_boot bootstrap statistics instead of a summary measure
partweight	A numeric vector specifying the weights of the corpus parts; if not specified, function returns unweighted estimate
directionality	Character string indicating the directionality of scaling. See details below. Possible values are "conventional" (default) and "gries"
formula	Character string indicating which formula to use for the calculation of DP. See details below. Possible values are "egbert_etal_2020" (default), "gries_2008", "lijffit_gries_2012".
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE

digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)
verbose	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE
print_score	Logical. Whether the dispersion score should be printed to the console; default is TRUE
suppress_warning	Logical. Whether warning messages should be suppressed; default is FALSE

Details

This function calculates weighted dispersion measures and bootstrap confidence intervals.

Author(s)

Lukas Soenning

See Also

disp_DP() for finer control over the calculation of DP

Examples

```
disp_DP_boot(
  subfreq = biber150_ice_gb[3,],
  partsize = biber150_ice_gb[1,],
  digits = 2,
  freq_adjust = TRUE,
  directionality = "conventional",
  formula = "gries_2008")
```

---

disp_DP_sboot *Stratified bootstrapping for Gries's* deviation of proportions

---

Description

This function implements stratified bootstrapping (and weighting) for Gries's dispersion measure DP (deviation of proportions). In addition to the full functionality offered by the function disp_DP(), it can be used to obtain weighted dispersion scores as well as bootstrap confidence intervals.

Usage

```
disp_DP_sboot(
  text_freq,
  text_size,
  corpus_parts = NULL,
  n_boot = 500,
```

```
  boot_ci = FALSE,
  conf_level = 0.95,
  return_distribution = FALSE,
  partweight = NULL,
  directionality = "conventional",
  formula = "egbert_etal_2020",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

text_freq	Integer value giving the frequency of the item in the text
text_size	Interger value giving the size (or length) of the text
corpus_parts	The corpus parts that form the basis of dispersion analysis; must be higher-level categories, above the text files
n_boot	Integer value specifying the number of bootstrap samples to draw; default: 500
boot_ci	Logical. Whether a percentile bootstrap confidence interval should be computed; default: FALSE
conf_level	Scalar giving the confidence level; default 0.95 for a 95% percentile CI
return_distribution	Logical. Whether the function should return a vector of all n_boot bootstrap statistics instead of a summary measure
partweight	A numeric vector specifying the weights of the corpus parts; if not specified, function returns unweighted estimate
directionality	Character string indicating the directionality of scaling. See details below. Possible values are "conventional" (default) and "gries"
formula	Character string indicating which formula to use for the calculation of DP. See details below. Possible values are "egbert_etal_2020" (default), "gries_2008", "lijffit_gries_2012".
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE
digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)

verbose — Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE

print_score — Logical. Whether the dispersion score should be printed to the console; default is TRUE

suppress_warning — Logical. Whether warning messages should be suppressed; default is FALSE

Details

This function performs stratified bootstrapping on dispersion measures. Stratified bootstrapping is used when the corpus parts represent text categories (e.g. genres, registers) that in turn consists of texts or text files. Since the resampling scheme implemented in bootstrapping should be as closely aligned with the data layout (and data-generation procedure) as closely as possible, resampling should not take place at the level of the text categories. Instead, it is the sampling units in corpus compilation – texts, text files, or speakers – that should be resampled. Stratified bootstrapping therefore respects the structure of the corpus and data.

Author(s)

Lukas Soenning

See Also

disp_DP() for finer control over the calculation of DP

Examples

```
disp_DP_sboot(
  text_freq = biber150_brown[87,],
  text_size = biber150_brown[1,],
  corpus_parts = as.character(metadata_brown$genre),
  digits = 2,
  freq_adjust = TRUE,
  directionality = "conventional",
  formula = "gries_2008")
```

---

disp_DP_tdm *Calculate Gries's* deviation of proportions *for a term-document matrix*

Description

This function calculates Gries's dispersion measure DP (deviation of proportions). It offers three different formulas and allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_DP_tdm(
  tdm,
  row_partsize = "first",
  directionality = "conventional",
  formula = "egbert_etal_2020",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  add_frequency = TRUE,
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_scores = TRUE
)
```

Arguments

tdm	A term-document matrix, where rows represent items and columns represent corpus parts; must also contain a row giving the size of the corpus parts (first or last row in the term-document matrix)
row_partsize	Character string indicating which row in the term-document matrix contains the size of the corpus parts. Possible values are "first" (default) and "last"
directionality	Character string indicating the directionality of scaling. See details below. Possible values are "conventional" (default) and "gries"
formula	Character string indicating which formula to use for the calculation of DP. See details below. Possible values are "egbert_etal_2020" (default), "gries_2008", "lijffit_gries_2012".
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"
add_frequency	Logical. Whether to add a column that gives the total number of occurrences of the item across a corpus parts; default is TRUE
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE
digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)
verbose	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE
print_scores	Logical. Whether the dispersion scores should be printed to the console; default is TRUE

Details

This function takes as input a term-document matrix and returns, for each item (i.e. each row) the dispersion measure DP. The rows in the input matrix represent the items, and the columns the corpus parts. Importantly, the term-document matrix must include an additional row that records the size of the corpus parts. For a proper term-document matrix, which includes all items that appear in the corpus, this can be added as a column margin, which sums the frequencies in each column. If the matrix only includes a selection of items drawn from the corpus, this information cannot be derived from the matrix and must be provided as a separate row.

- Directionality: DP ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option.
- Formula: Irrespective of the directionality of scaling, four formulas for DP exist in the literature (see below for details). This is because the original version proposed by Gries (2008: 415), which is commonly denoted as DP (and here referenced by the value `"gries_2008"`) does not always reach its theoretical limits of 0 and 1. For this reason, modifications have been suggested, starting with Gries (2008: 419) himself, who referred to this version as $DPnorm$. This version is not implemented in the current package, because Lijffit & Gries (2012) updated $DPnorm$ to ensure that it also works as intended when corpus parts differ in size; this version is represented by the value `"lijffit_gries_2012"` and often denoted using subscript notation DP_{norm} . Finally, Egbert et al. (2020: 99) suggest a further modification to ensure proper behavior in settings where the item occurs in only one corpus part. They label this version D_P. In the current function, it is the default and represented by the value `"egbert_etal_2020"`.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.
  - To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).
  - To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()`.

In the formulas given below, the following notation is used:

- k the number of corpus parts
- t_i a proportional quantity; the subfrequency in part i divided by the total number of occurrences of the item in the corpus (i.e. the sum of all subfrequencies)
- w_i a proportional quantity; the size of corpus part i divided by the size of the corpus (i.e. the sum of the part sizes)

The value `"gries_2008"` implements the original version proposed by Gries (2008: 415). Note that while the following formula represents Gries scaling (0 = even, 1 = uneven), in the current function the directionality is controlled separately using the argument `directionality`.

$\frac{\sum_i^k |t_i - w_i|}{2}$ (Gries 2008)

The value `"lijffit_gries_2012"` implements the modified version described by Lijffit & Gries (2012). Again, the following formula represents Gries scaling (0 = even, 1 = uneven), but the directionality is handled separately in the current function. The notation $min\{w_i\}$ refers to the w_i value of the smallest corpus part.

$\frac{\sum_i^k |t_i - w_i|}{2} \times \frac{1}{1 - min\{w_i\}}$ (Lijffijt & Gries 2012)

The value `"egbert_etal_2020"` (default) selects the modification suggested by Egbert et al. (2020: 99). The following formula represents conventional scaling (0 = uneven, 1 = even). The notation $min\{w_i : t_i > 0\}$ refers to the w_i value among those corpus parts that include at least one occurrence of the item.

$1 - \frac{\sum_i^k |t_i - w_i|}{2} \times \frac{1}{1 - min\{w_i : t_i > 0\}}$ (Egbert et al. 2020)

Value

A data frame with one row per item

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115. doi:10.1075/ijcl.18010.egb

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Lijffijt, Jefrey & Stefan Th. Gries. 2012. Correction to Stefan Th. Gries' 'Dispersions and adjusted frequencies in corpora'. *International Journal of Corpus Linguistics* 17(1). 147–149. doi:10.1075/ijcl.17.1.08lij

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
disp_DP_tdm(
  tdm = biber150_spokenBNC2014[1:20,],
  row_partsize = "first",
  directionality = "conventional",
  formula = "gries_2008",
  freq_adjust = FALSE)
```

`disp_R` *Calculate the dispersion measure 'range'*

Description

This function calculates the dispersion measure 'range'. It offers three different versions: 'absolute range' (the number of corpus parts containing at least one occurrence of the item), 'relative range' (the proportion of corpus parts containing at least one occurrence of the item), and 'relative range with size' (relative range that takes into account the size of the corpus parts). The function also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_R(
  subfreq,
  partsize,
  type = "relative",
  freq_adjust = FALSE,
  freq_adjust_method = "pervasive",
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

`subfreq`	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
`partsize`	A numeric vector specifying the size of the corpus parts

`type` Character string indicating which type of range to calculate. See details below. Possible values are `"relative"` (default), `"absolute"`, `"relative_withsize"`

`freq_adjust` Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is `FALSE`

`freq_adjust_method` Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are `"pervasive"` (default) and `"even"`

`unit_interval` Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is `TRUE`

`digits` Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)

`verbose` Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is `TRUE`

`print_score` Logical. Whether the dispersion score should be printed to the console; default is `TRUE`

`suppress_warning` Logical. Whether warning messages should be suppressed; default is `FALSE`

Details

The function calculates the dispersion measure 'range' based on a set of subfrequencies (number of occurrences of the item in each corpus part) and a matching set of part sizes (the size of the corpus parts, i.e. number of word tokens). Three different types of range measures can be calculated:

- Absolute range: The number of corpus parts containing at least one occurrence of the item
- Relative range: The proportion of corpus parts containing at least one occurrence of the item; this version of 'range' follows the conventional scaling of dispersion measures (1 = widely dispersed)
- Relative range with size (see Gries 2022: 179-180; Gries 2024: 27-28): Relative range that takes into account the size of the corpus parts. Each corpus part contributes to this version of range in proportion to its size. Suppose there are 100 corpus parts, and part 1 is relatively short, accounting for 1/200 of the words in the whole corpus. If the item occurs in part 1, ordinary relative range increases by 1/100, since each part receives the same weight. Relative range with size, on the other hand, increases by 1/200, i.e. the relative size of the corpus part; this version of range weights corpus parts proportionate to their size.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses

a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.

- To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).
- To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()`.

Value

A numeric value

Author(s)

Lukas Soenning

References

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Examples

```
disp_R(
  subfreq = c(0, 0, 1, 2, 5),
  partsize = rep(1000, 5),
  type = "relative",
  freq_adjust = FALSE)
```

`disp_R_tdm` *Calculate the dispersion measure 'range' for a term-document matrix*

Description

This function calculates the dispersion measure 'range'. It offers three different versions: 'absolute range' (the number of corpus parts containing at least one occurrence of the item), 'relative range' (the proportion of corpus parts containing at least one occurrence of the item), and 'relative range with size' (relative range that takes into account the size of the corpus parts). The function also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_R_tdm(
  tdm,
  row_partsize = "first",
  type = "relative",
  freq_adjust = FALSE,
  freq_adjust_method = "pervasive",
  add_frequency = TRUE,
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_scores = TRUE
)
```

Arguments

tdm	A term-document matrix, where rows represent items and columns represent corpus parts; must also contain a row giving the size of the corpus parts (first or last row in the term-document matrix)
row_partsize	Character string indicating which row in the term-document matrix contains the size of the corpus parts. Possible values are "first" (default) and "last"
type	Character string indicating which type of range to calculate. See details below. Possible values are "relative" (default), "absolute", "relative_withsize"
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "pervasive" (default) and "even"
add_frequency	Logical. Whether to add a column that gives the total number of occurrences of the item across a corpus parts; default is TRUE
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE
digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)
verbose	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE
print_scores	Logical. Whether the dispersion scores should be printed to the console; default is TRUE

Details

This function takes as input a term-document matrix and returns, for each item (i.e. each row) the dispersion measure 'range'. The rows in the input matrix represent the items, and the columns the corpus parts. Importantly, the term-document matrix must include an additional row that records the size of the corpus parts. For a proper term-document matrix, which includes all items that appear

in the corpus, this can be added as a column margin, which sums the frequencies in each column. If the matrix only includes a selection of items drawn from the corpus, this information cannot be derived from the matrix and must be provided as a separate row.

Three different types of range measures can be calculated:

- Absolute range: The number of corpus parts containing at least one occurrence of the item
- Relative range: The proportion of corpus parts containing at least one occurrence of the item; this version of 'range' follows the conventional scaling of dispersion measures (1 = widely dispersed)
- Relative range with size (see Gries 2022: 179-180; Gries 2024: 27-28): Relative range that takes into account the size of the corpus parts. Each corpus part contributes to this version of range in proportion to its size. Suppose there are 100 corpus parts, and part 1 is relatively short, accounting for 1/200 of the words in the whole corpus. If the item occurs in part 1, ordinary relative range increases by 1/100, since each part receives the same weight. Relative range with size, on the other hand, increases by 1/200, i.e. the relative size of the corpus part; this version of range weights corpus parts proportionate to their size.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `"even"`. For details and explanations, see `vignette("frequency-adjustment")`.
  - To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).
  - To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()`.

Value

A data frame with one row per item

Author(s)

Lukas Soenning

References

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Examples

```
disp_R_tdm(
  tdm = biber150_spokenBNC2014[1:20,],
  row_partsize = "first",
  type = "relative",
  freq_adjust = FALSE)
```

---

disp_S *Calculate the dispersion measure S*

Description

This function calculates the dispersion measure S (Rosengren 1971) and allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_S(
  subfreq,
  partsize,
  directionality = "conventional",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_score = TRUE,
  suppress_warning = FALSE
)
```

Arguments

`subfreq`	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
`partsize`	A numeric vector specifying the size of the corpus parts
`directionality`	Character string indicating the directionality of scaling. See details below. Possible values are `"conventional"` (default) and `"gries"`

`freq_adjust` Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is `FALSE`

`freq_adjust_method`
Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are `"even"` (default) and `"pervasive"`

`unit_interval` Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is `TRUE`

`digits` Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)

`verbose` Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is `TRUE`

`print_score` Logical. Whether the dispersion score should be printed to the console; default is `TRUE`

`suppress_warning`
Logical. Whether warning messages should be suppressed; default is `FALSE`

Details

The function calculates the dispersion measure S based on a set of subfrequencies (number of occurrences of the item in each corpus part) and a matching set of part sizes (the size of the corpus parts, i.e. number of word tokens).

- Directionality: S ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed here in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.
  - To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).

  - To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()` and `vignette("frequency-adjustment")`.

In the formulas given below, the following notation is used:

- k the number of corpus parts
- T_i the absolute subfrequency in part i
- w_i a proportional quantity; the size of corpus part i divided by the size of the corpus (i.e. the sum of the part sizes)

S is the dispersion measure proposed by Rosengren (1971); the formula uses conventional scaling: $\frac{(\sum_i^k r_i \sqrt{w_i T_i}}{N}$

Value

A numeric value

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
disp_S(
  subfreq = c(0,0,1,2,5),
  partsize = rep(1000, 5),
  directionality = "conventional")
```

disp_S_tdm *Calculate the dispersion measure S for a term-document matrix*

Description

This function calculates the dispersion measure S (Rosengren 1971) and allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_S_tdm(
  tdm,
  row_partsize = "first",
  directionality = "conventional",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  unit_interval = TRUE,
  add_frequency = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_scores = TRUE
)
```

Arguments

tdm	A term-document matrix, where rows represent items and columns represent corpus parts; must also contain a row giving the size of the corpus parts (first or last row in the term-document matrix)
row_partsize	Character string indicating which row in the term-document matrix contains the size of the corpus parts. Possible values are "first" (default) and "last"
directionality	Character string indicating the directionality of scaling. See details below. Possible values are "conventional" (default) and "gries"
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE
add_frequency	Logical. Whether to add a column that gives the total number of occurrences of the item across a corpus parts; default is TRUE
digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)

verbose	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE
print_scores	Logical. Whether the dispersion scores should be printed to the console; default is TRUE

Details

This function takes as input a term-document matrix and returns, for each item (i.e. each row) the dispersion measure S. The rows in the input matrix represent the items, and the columns the corpus parts. Importantly, the term-document matrix must include an additional row that records the size of the corpus parts. For a proper term-document matrix, which includes all items that appear in the corpus, this can be added as a column margin, which sums the frequencies in each column. If the matrix only includes a selection of items drawn from the corpus, this information cannot be derived from the matrix and must be provided as a separate row.

- Directionality: S ranges from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. This is the default. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = 'gries'` to choose this option.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`pervasive`) or evenness (`even`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.
  - To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`pervasive`), or they are assigned to the smallest corpus part(s) (`even`).
  - To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`pervasive`), or they are allocated to corpus parts in proportion to their size (`even`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()`.

In the formulas given below, the following notation is used:

- k the number of corpus parts
- T_i the absolute subfrequency in part i
- w_i a proportional quantity; the size of corpus part i divided by the size of the corpus (i.e. the sum of the part sizes)

S is the dispersion measure proposed by Rosengren (1971); the formula uses conventional scaling: $\frac{(\sum_i^k r_i \sqrt{w_i T_i}}{N}$

Value

A data frame with one row per item

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
disp_S_tdm(
  tdm = biber150_spokenBNC2014[1:20,],
  row_partsize = "first",
  directionality = "conventional")
```

`disp_tdm` *Calculate parts-based dispersion measures for a term-document matrix*

Description

This function calculates a number of parts-based dispersion measures and allows the user to choose the directionality of scaling, i.e. whether higher values denote a more even or a less even distribution. It also offers the option of calculating frequency-adjusted dispersion scores.

Usage

```
disp_tdm(
  tdm,
  row_partsize = "first",
  directionality = "conventional",
  freq_adjust = FALSE,
  freq_adjust_method = "even",
  add_frequency = TRUE,
  unit_interval = TRUE,
  digits = NULL,
  verbose = TRUE,
  print_scores = TRUE,
  suppress_warning = FALSE
)
```

Arguments

tdm	A term-document matrix, where rows represent items and columns represent corpus parts; must also contain a row giving the size of the corpus parts (first or last row in the term-document matrix)
row_partsize	Character string indicating which row in the term-document matrix contains the size of the corpus parts. Possible values are "first" (default) and "last"
directionality	Character string indicating the directionality of scaling. See details below. Possible values are "conventional" (default) and "gries"
freq_adjust	Logical. Whether dispersion score should be adjusted for frequency (i.e. whether frequency should be 'partialed out'); default is FALSE
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"
add_frequency	Logical. Whether to add a column that gives the total number of occurrences of the item across a corpus parts; default is TRUE
unit_interval	Logical. Whether frequency-adjusted scores that exceed the limits of the unit interval should be replaced by 0 and 1; default is TRUE
digits	Rounding: Integer value specifying the number of decimal places to retain (default: no rounding)
verbose	Logical. Whether additional information (on directionality, formulas, frequency adjustment) should be printed; default is TRUE
print_scores	Logical. Whether the dispersion scores should be printed to the console; default is TRUE
suppress_warning	Logical. Whether warning messages should be suppressed; default is FALSE

Details

This function takes as input a term-document matrix and returns, for each item (i.e. each row) a variety of dispersion measures. The rows in the input matrix represent the items, and the columns

the corpus parts. Importantly, the term-document matrix must include an additional row that records the size of the corpus parts. For a proper term-document matrix, which includes all items that appear in the corpus, this can be added as a column margin, which sums the frequencies in each column. If the matrix only includes a selection of items drawn from the corpus, this information cannot be derived from the matrix and must be provided as a separate row.

- Directionality: The scores for all measures range from 0 to 1. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; this is implemented by the value `gries`.
- Frequency adjustment: Dispersion scores can be adjusted for frequency using the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208). The frequency-adjusted score for an item considers the lowest and highest possible level of dispersion it can obtain given its overall corpus frequency as well as the number (and size) of corpus parts. The unadjusted score is then expressed relative to these endpoints, where the dispersion minimum is set to 0, and the dispersion maximum to 1 (expressed in terms of conventional scaling). The frequency-adjusted score falls between these bounds and expresses how close the observed distribution is to the theoretical maximum and minimum. This adjustment therefore requires a maximally and a minimally dispersed distribution of the item across the parts. These hypothetical extremes can be built in different ways. The method used by Gries (2022, 2024) uses a computationally expensive procedure that finds the distribution that produces the highest value on the dispersion measure of interest. The current function constructs extreme distributions in a different way, based on the distributional features pervasiveness (`"pervasive"`) or evenness (`"even"`). You can choose between these with the argument `freq_adjust_method`; the default is `even`. For details and explanations, see `vignette("frequency-adjustment")`.
  - To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`"pervasive"`), or they are assigned to the smallest corpus part(s) (`"even"`).
  - To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`"pervasive"`), or they are allocated to corpus parts in proportion to their size (`"even"`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. See documentation for `find_max_disp()` and `vignette("frequency-adjustment")`.

The following measures are computed, listed in chronological order (see details below):

- R_{rel} (Keniston 1920)
- D (Juilland & Chang-Rodriguez 1964)
- D_2 (Carroll 1970)
- S (Rosengren 1971)
- D_P (Gries 2008; modification: Egbert et al. 2020)
- D_A (Burch et al. 2017)
- D_{KL} (Gries 2024)

In the formulas given below, the following notation is used:

- k the number of corpus parts

- T_i the absolute subfrequency in part i
- t_i a proportional quantity; the subfrequency in part i divided by the total number of occurrences of the item in the corpus (i.e. the sum of all subfrequencies)
- W_i the absolute size of corpus part i
- w_i a proportional quantity; the size of corpus part i divided by the size of the corpus (i.e. the sum of the part sizes)
- R_i the normalized subfrequency in part i, i.e. the subfrequency divided by the size of the corpus part
- r_i a proportional quantity; the normalized subfrequency in part i divided by the sum of all normalized subfrequencies
- N corpus frequency, i.e. the total number of occurrence of the item in the corpus

Note that the formulas cited below differ in their scaling, i.e. whether 1 reflects an even or an uneven distribution. In the current function, this behavior is overridden by the argument `directionality`. The specific scaling used in the formulas below is therefore irrelevant.

R_{rel} refers to the relative range, i.e. the proportion of corpus parts containing at least one occurrence of the item

D denotes Juilland's D and is calculated as follows (this formula uses conventional scaling); $\bar{R}_i$ denotes the average over the normalized subfrequencies:

$$1 - \sqrt{\frac{\sum_{i=1}^{k}(R_i - \bar{R}_i)^2}{k}} \times \frac{1}{\bar{R}_i\sqrt{k-1}}$$

D_2 denotes the index proposed by Carroll (1970); the following formula uses conventional scaling:

$$\frac{\sum_i^k r_i \log_2 \frac{1}{r_i}}{\log_2 k}$$

S is the dispersion measure proposed by Rosengren (1971); the formula uses conventional scaling:

$$\frac{(\sum_i^k r_i\sqrt{w_i T_i}}{N}$$

D_P represents Gries's *deviation of proportions*; the following formula is the modified version suggested by Egbert et al. (2020: 99); it implements conventional scaling (0 = uneven, 1 = even) and the notation $min\{w_i : t_i > 0\}$ refers to the w_i value among those corpus parts that include at least one occurrence of the item.

$$1 - \frac{\sum_i^k |t_i - w_i|}{2} \times \frac{1}{1 - min\{w_i : t_i > 0\}}$$

D_A is a measure introduced into dispersion analysis by Burch et al. (2017). The following formula is the one used by Egbert et al. (2020: 98); it relies on normalized frequencies and therefore works with corpus parts of different size. The formula represents conventional scaling (0 = uneven, 1 = even):

$$1 - \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}|R_i - R_j|}{\frac{k(k-1)}{2}} \times \frac{1}{2\frac{\sum_i^k R_i}{k}}$$

The current function uses a formula that may be found in Wilcox (1973: 343). It relies on the proportional r_i values instead of the normalized subfrequencies R_i:

$$1 - \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}|r_i - r_j|}{k-1}$$

Since this formula is computationally expensive, the function actually uses the computational shortcut given in Wilcox (1973: 343). Critically, the proportional quantities r_i must first be sorted in

decreasing order. Only after this rearrangement can the shortcut version be applied. We will refer to this rearranged version of r_i as r_i^{sorted}:

$\frac{2\left(\sum_{i=1}^{k}(i \times r_i^{sorted})-1\right)}{k-1}$ (Wilcox 1973: 343)

D_{KL} denotes a measure proposed by Gries (2020, 2021); for standardization, it uses the odds-to-probability transformation (Gries 2024: 90) and represents Gries scaling (0 = even, 1 = uneven):

$\frac{\sum_i^k t_i \log_2 \frac{t_i}{w_i}}{1+\sum_i^k t_i \log_2 \frac{t_i}{w_i}}$

Value

A data frame with one row per item

Author(s)

Lukas Soenning

References

Burch, Brent, Jesse Egbert & Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216. doi:10.1558/jrds.33066

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115. doi:10.1075/ijcl.18010.egb

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 99–118. New York: Springer. doi:10.1007/9783030-462161_5

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2). 1–33. doi:10.32714/ricl.09.02.02

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Keniston, Hayward. 1920. Common words in Spanish. *Hispania* 3(2). 85–96. doi:10.2307/331305

Lijffijt, Jefrey & Stefan Th. Gries. 2012. Correction to Stefan Th. Gries' 'Dispersions and adjusted frequencies in corpora'. *International Journal of Corpus Linguistics* 17(1). 147–149. doi:10.1075/ijcl.17.1.08lij

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

See Also

For finer control over the calculation of several dispersion measures:

- `disp_R_tdm()` for $Range$
- `disp_DP_tdm()` for D_P
- `disp_DA_tdm()` for D_A
- `disp_DKL_tdm()` for D_{KL}

Examples

```
disp_tdm(
  tdm = biber150_spokenBNC2014[1:20,],
  row_partsize = "first",
  directionality = "conventional",
  freq_adjust = FALSE)
```

---

find_max_disp *Find the maximally dispersed distribution of an item across corpus parts*

Description

This function returns the (hypothetical) distribution of subfrequencies that represents the highest possible level of dispersion for a given item across a particular set of corpus parts. It requires a vector of subfrequencies and a vector of corpus part sizes. This distribution is required for the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208) to obtain frequency-adjusted dispersion scores.

Usage

```
find_max_disp(subfreq, partsize, freq_adjust_method = "even")
```

Arguments

`subfreq`	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
`partsize`	A numeric vector specifying the size of the corpus parts
`freq_adjust_method`	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are `"even"` (default) and `"pervasive"`

Details

This function creates a hypothetical distribution of the total number of occurrences of the item (i.e. the sum of its subfrequencies) across corpus parts. To obtain the highest possible level of dispersion, the argument `freq_adjust_method` allows the user to choose between two distributional features: pervasiveness (`pervasive`) or evenness (`even`). For details and explanations, see `vignette("frequency-adjustment")`. To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`pervasive`), or they are allocated to corpus parts in proportion to their size (`even`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. Since the dispersion of an item that occurs only once in the corpus (hapaxes) cannot be sensibly measured or manipulated, such items are disregarded; the function returns their observed subfrequencies.

Value

An integer vector the same length as `partsize`

Author(s)

Lukas Soenning

References

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Examples

```
find_max_disp(
  subfreq = c(0,0,1,2,5),
  partsize = c(100, 100, 100, 500, 1000),
  freq_adjust_method = "pervasive")
```

---

find_max_disp_tdm *Find the maximally dispersed distribution of each item in a term-document matrix*

Description

This function takes as input a term-document matrix and returns, for each item (i.e. row), the (hypothetical) distribution of subfrequencies that represents the highest possible level of dispersion for the item across the corpus parts. This distribution is required for the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208) to obtain frequency-adjusted dispersion scores.

Usage

```
find_max_disp_tdm(
  tdm,
  row_partsize = "first",
  freq_adjust_method = freq_adjust_method
)
```

Arguments

`tdm`	A term-document matrix, where rows represent items and columns represent corpus parts; must also contain a row giving the size of the corpus parts (first or last row in the term-document matrix)
`row_partsize`	Character string indicating which row in the term-document matrix contains the size of the corpus parts. Possible values are `"first"` (default) and `"last"`
`freq_adjust_method`	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are `"even"` (default) and `"pervasive"`

Details

This function takes as input a term-document matrix and creates, for each item in the matrix, a hypothetical distribution of the total number of occurrences of the item (i.e. the sum of the subfrequencies) across corpus parts. To obtain the highest possible level of dispersion, the argument `freq_adjust_method` allows the user to choose between two distributional features: pervasiveness (`pervasive`) or evenness (`even`). For details and explanations, see `vignette("frequency-adjustment")`. To obtain the highest possible level of dispersion, the occurrences are either spread as broadly across corpus parts as possible (`pervasive`), or they are allocated to corpus parts in proportion to their size (`even`). The choice between these methods is particularly relevant if corpus parts differ considerably in size. Since the dispersion of items that occur only once in the corpus (hapaxes) cannot be sensibly measured or manipulated, such items are disregarded; the function returns their observed subfrequencies.

Value

A matrix of integers with one row per item and one column per corpus part

Author(s)

Lukas Soenning

References

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins.

See Also

find_max_disp()

Examples

```
find_max_disp_tdm(
  tdm = biber150_spokenBNC2014[1:10,],
  row_partsize = "first",
  freq_adjust_method = "even")
```

---

find_min_disp *Find the minimally dispersed distribution of an item across corpus parts*

Description

This function returns the (hypothetical) distribution of subfrequencies that represents the smallest possible level of dispersion for a given item across a particular set of corpus parts. It requires a vector of subfrequencies and a vector of corpus part sizes. This distribution is required for the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208) to obtain frequency-adjusted dispersion scores.

Usage

```
find_min_disp(subfreq, partsize, freq_adjust_method = "even")
```

Arguments

subfreq	A numeric vector of subfrequencies, i.e. the number of occurrences of the item in each corpus part
partsize	A numeric vector specifying the size of the corpus parts
freq_adjust_method	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are "even" (default) and "pervasive"

Details

This function creates a hypothetical distribution of the total number of occurrences of the item (i.e. the sum of its subfrequencies) across corpus parts. To obtain the lowest possible level of dispersion, the argument freq_adjust_method allows the user to choose between two distributional features: pervasiveness (pervasive) or evenness (even). For details and explanations, see vignette("frequency-adjustment"). To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (pervasiveness), or they are assigned to the smallest corpus part(s) (even). Since the dispersion of items that occur only once in the corpus (hapaxes) cannot be sensibly measured or manipulated, such items are disregarded; the function returns their observed subfrequencies. The function reuses code segments from Gries's (2025) 'KLD4C' package (from the function most.uneven.distr()).

Value

An integer vector the same length as `partsize`

Author(s)

Lukas Soenning

References

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Gries, Stefan Th. 2025. *KLD4C: Gries 2024: Tupleization of corpus linguistics*. R package version 1.01. (available from https://www.stgries.info/research/kld4c/kld4c.html)

Examples

```
find_min_disp(
  subfreq = c(0,0,1,2,5),
  partsize = rep(1000, 5),
  freq_adjust_method = "even")
```

---

find_min_disp_tdm *Find the minimally dispersed distribution of each item in a term-document matrix*

Description

This function takes as input a term-document matrix and returns, for each item (i.e. row), the (hypothetical) distribution of subfrequencies that represents the smallest possible level of dispersion for the item across the corpus parts. This distribution is required for the min-max transformation proposed by Gries (2022: 184-191; 2024: 196-208) to obtain frequency-adjusted dispersion scores.

Usage

```
find_min_disp_tdm(
  tdm,
  row_partsize = "first",
  freq_adjust_method = freq_adjust_method
)
```

Arguments

`tdm`	A term-document matrix, where rows represent items and columns represent corpus parts; must also contain a row giving the size of the corpus parts (first or last row in the term-document matrix)
`row_partsize`	Character string indicating which row in the term-document matrix contains the size of the corpus parts. Possible values are `"first"` (default) and `"last"`
`freq_adjust_method`	Character string indicating which method to use for devising dispersion extremes. See details below. Possible values are `"even"` (default) and `"pervasive"`

Details

This function takes as input a term-document matrix and creates, for each item in the matrix, a hypothetical distribution of the total number of occurrences of the item (i.e. the sum of the subfrequencies) across corpus parts. To obtain the lowest possible level of dispersion, the argument `freq_adjust_method` allows the user to choose between two distributional features: pervasiveness (`pervasive`) or evenness (`even`). For details and explanations, see `vignette("frequency-adjustment")`. To obtain the lowest possible level of dispersion, the occurrences are either allocated to as few corpus parts as possible (`pervasiveness`), or they are assigned to the smallest corpus part(s) (`even`). Since the dispersion of an item that occurs only once in the corpus (hapaxes) cannot be sensibly measured or manipulated, such items are disregarded; the function returns their observed subfrequencies. The function reuses code segments from Gries's (2025) 'KLD4C' package (from the function `most.uneven.distr()`).

Value

A matrix of integers with one row per item and one column per corpus part

Author(s)

Lukas Soenning

References

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205. doi:10.1075/jsls.21029.gri

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: Benjamins. doi:10.1075/scl.115

Gries, Stefan Th. 2025. *KLD4C: Gries 2024: Tupleization of corpus linguistics*. R package version 1.01. (available from https://www.stgries.info/research/kld4c/kld4c.html)

See Also

`find_min_disp()`

Examples

```
find_min_disp_tdm(
  tdm = biber150_spokenBNC2014[1:10,],
  row_partsize = "first",
  freq_adjust_method = "even")
```

`fpower` *Re-express proportions using the folded power transformation*

Description

This function takes as input a vector of proportions (or, more generally, scores in the unit interval [0,1]) and re-expresses them using Tukey's folded power transformation. It allows the user to decide whether the transformed scores should be mapped to the [-1, +1] interval (default), or whether they may extend beyond these limits.

Usage

```
fpower(x, lambda = 1, scaling = "plus_minus_1")
```

Arguments

`x`	A numeric vector of scores in the unit interval [0,1]; 0 and 1 are allowed but throw an error message when lambda = 0
`lambda`	Numeric value of the power transformation, which can range between 0 (limiting case: logit transformation) and 1 (no transformation)
`scaling`	Character string indicating whether scores should be re-expressed to the [-1, 1] interval (`plus_minus_1`) or allowed to stretch beyond these limits (`free`)

Details

This function allows the user to apply a variety of folded power transformations to quantities bounded between 0 and 1. Different values may be specified for the power of the transformation (λ `lambda`), but only powers between 0 and 1 are supported. Two versions of the folded power transformation are available. The first maps transformed values to the [-1, +1] interval:

$x^\lambda - (1-x)^\lambda$

The second version does not impose these limits:

$(x^\lambda - (1-x)^\lambda)/\lambda$

For `lambda` equal to 0, the logit transformation is implemented as a limiting case; note that input scores of 0 and 1 are not allowed when `lambda` is set to 0 `lambda` = 0.14 gives a close approximation to the probit transformation (see Fox 2016: 74) while accepting input score of 0 and 1 `lambda` = 1/3 implements folded cube roots `lambda` = 0.41 gives a close approximation to the arcsine-square-root (or angular) transformation (see Fox 2016: 74) `lambda` = 0.5 implements folded roots

This function was written with the help of ChatGPT (version GPT-5.1; OpenAI 2025)

Value

A numeric vector

Author(s)

Lukas Soenning

References

OpenAI. (2025). ChatGPT (GPT-5.1) Large language model. https://chat.openai.com

Examples

```
fpower(
  seq(0, 1, .1),
  lambda = .14,
  scaling = "plus_minus_1")
```

---

fpower_trans *Tukey's folded power transformation*

Description

Tukey's folded power transformation

Usage

```
fpower_trans(lambda = 0)
```

Arguments

lambda	Numeric value of the applied power transformation, which can range between 0 (limiting case: logit transformation) and 1 (no transformation)

Details

This function was written with the help of ChatGPT (version GPT-5.1; OpenAI 2025)

Value

A numeric vector

References

OpenAI. (2025). ChatGPT (GPT-5.1) Large language model. https://chat.openai.com

Examples

```
plot(fpower_trans(lambda = .5), xlim = c(0, 1))
```

`invfpower` *Back-transform folded-power-transformed scores to the unit interval [0,1]*

Description

This function takes as input a vector of transformed scores, i.e. values that were originally in the unit interval [0, 1] but which were re-expressed using Tukey's folded power transformation. It allows back-transformation of two versions of folded powers: Those that are mapped to the [-1, +1] interval and those that aren't.

Usage

```
invfpower(y, lambda = 1, scaling = "plus_minus_1")
```

Arguments

`y`	A numeric vector of folded-power-transformed scores
`lambda`	Numeric value of the applied power transformation, which can range between 0 (limiting case: logit transformation) and 1 (no transformation)
`scaling`	Character string indicating whether scores were re-expressed to the [-1, 1] interval (`plus_minus_1`) or not (`free`)

Details

This function was written with the help of ChatGPT (version GPT-5.1; OpenAI 2025)

Value

A numeric vector

Author(s)

Lukas Soenning

References

OpenAI. (2025). ChatGPT (GPT-5.1) Large language model. https://chat.openai.com

Examples

```
invfpower(
  seq(-1, 1, .1),
  lambda = .14,
  scaling = "plus_minus_1")
```

metadata_brown *Text metadata for the Brown corpus*

Description

This dataset provides metadata for the text files in the Brown corpus (Francis & Kučera 1979). It maps standardized file names to the textual categories macro genre and genre, and records the length of each text file (in the total number of word and nonword tokens). Macro genres and genres are ordered based on the sampling frame informing the design of the Brown family of corpora (see https://listings.lib.msu.edu/public-corpora/cd421/manuals/brown/INDEX.HTM).

Usage

```
metadata_brown
```

Format

`metadata_brown`**:**

A data frame with 500 rows and 4 columns:

text_file Standardized name of the text file (e.g. "A01", "J58", "R07")

macro_genre 4 macro genres ("press", "general_prose", "learned", "fiction"); ordered factor

genre 15 genres (e.g. "press_editorial", "popular_lore", "adventure_western_fiction"); ordered factor

word_count The length of the text file, expressed as the number of (word and nonword) tokens

Source

Francis, W. Nelson & Henry Kučera. 1979. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers (Brown)*. Providence, RI: Brown University.

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics*. Cambridge: Cambridge University Press.

metadata_ice_gb *Text metadata for ICE-GB*

Description

This dataset provides metadata for the text files in ICE-GB (Nelson et al. 2002). It maps standardized file names to various textual categories such as mode of production, macro genre and genre, and records the length of each text file (in the total number of word and nonword tokens). Text categories, macro genres and genres are ordered based on the sampling frame informing the design of the ICE family of corpora (see https://www.ice-corpora.uzh.ch/en/design.html).

Usage

```
metadata_ice_gb
```

Format

`metadata_ice_gb`**:**

A data frame with 500 rows and 7 columns:

text_file Standardized name of the text file (e.g. "s1a-001", "w1b-008", "w2d-018")

mode Mode of production ("spoken" vs. "written")

text_category 4 higher-level text categories ("dialogues", "monologues", "non-printed", "printed"); ordered factor

macro_genre 12 macro genres (e.g. "private_dialogues", "student_writing", "reportage"); ordered factor

genre 32 genres (e.g. "phonecalls", "unscripted_speeches", "novels_short_stories"); ordered factor

genre_short Short label for the genre (see Schützler 2023: 228); ordered factor

word_count The length of the text file, expressed as the number of (word and nonword) tokens

Source

https://www.ice-corpora.uzh.ch/en/design.html

Greenbaum, Sidney. 1996. Introducing ICE. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 3–12. Clarendon Press.

Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. John Benjamins.

Schützler, Ole. 2023. *Concessive constructions in varieties of English*. Language Science Press. doi:10.5281/zenodo.8375010

`metadata_spokenBNC1994`

Speaker metadata for the Spoken BNC1994

Description

This dataset provides some metadata for speakers in the demographically sampled part of the Spoken BNC1994 (Crowdy 1995), including information on age, gender, and the total number of word tokens contributed to the corpus.

Usage

```
metadata_spokenBNC1994
```

Format

`metadata_spokenBNC1994`**:**

A data frame with 1,017 rows and 7 columns:

speaker_id Speaker ID (e.g. "PS002", "PS003")

age_group Age group, based on the BNC1994 scheme ("0-14", "15-24", "25-34", "35-44", "45-59", "60+", "Unknown")

gender Speaker gender ("Female" vs. "Male")

age Age of speaker; if actual age is not available, imputed based on `age_group` and `age_bin`

n_tokens Number of word tokens the speaker contributed to the corpus

age_bin Age group, based on the BNC2014 scheme ("0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70+")

Source

Crowdy, Steve. 1995. The BNC spoken corpus. In Geoffrey Leech, Greg Myers & Jenny Thomas (eds.), *Spoken English on Computer: Transcription, Mark-Up and Annotation*, 224–234. Harlow: Longman.

`metadata_spokenBNC2014`

Speaker metadata for the Spoken BNC2014

Description

This dataset provides some metadata for the speakers in the Spoken BNC2014 (Love et al. 2017), including information on age, gender, and the total number of word tokens contributed to the corpus.

Usage

```
metadata_spokenBNC2014
```

Format

`metadata_spokenBNC2014`**:**

A data frame with 668 rows and 6 columns:

speaker_id Speaker ID (e.g. "S0001", "S0002")

age_group Age group, based on the BNC1994 scheme ("0-14", "15-24", "25-34", "35-44", "45-59", "60+", "Unknown")

gender Speaker gender ("Female" vs. "Male")

age Age of speaker; if actual age is not available, imputed based on `age_group` and `age_bin`

n_tokens Number of word tokens the speaker contributed to the corpus

age_bin Age group, based on the BNC2014 scheme ("0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70+")

Source

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319–344.

`scale_x_dispersion` *Add annotation to tick marks on x-axis (ggplot2) to clarify the directionality of scaling for dispersion scores*

Description

This function can be used when plotting dispersion scores with ggplot2. It forces the x-axis to extend from 0 to 1 and adds verbal information at the endpoints to clarify the directionality of scaling. For `conventional` scaling, these are "(uneven) 0" and "(even) 1", for `gries` scaling, these are "(even) 0" and "(uneven) 1".

Usage

```
scale_x_dispersion(directionality, n_breaks = 5, leading_zero = TRUE, ...)
```

Arguments

`directionality`	Character string indicating the directionality of scaling. Must match the way the dispersion scores were calculated. See details below. Possible values are `"conventional"` and `"gries"`
`n_breaks`	Number of major scale breaks: Integer value specifying the number of tick marks to display (default: `5`)
`leading_zero`	Logical. Whether the tick mark labels should include a leading 0 ("0.50") or not (".50"); default is `TRUE`
`...`	Other arguments passed on to `scale_x_continuous()`

Details

This function modifies the x-axis in a ggplot2 object. It forces the axis to extend from 0 to 1 and adds the labels "(even)" and "(uneven)" at the endpoints of the scale (0 and 1), to make clear which value (0 or 1) denotes a maximally even/dispersed/balanced distribution of subfrequencies across corpus parts. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. In the `{tlda}` package, this is the default setting for all measures. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option. The function implements no default, so the user must specify which directionality was used when calculating the scores.

Value

The ggplot2 function scale_x_continuous with the appropriate settings for the argument `limits`, `breaks`, and `labels`.

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
if (require("ggplot2")) {
  ggplot(
    data = data.frame(
      dispersion = stats::runif(100, 0, 1)),
    aes(x = dispersion)) +
    geom_dotplot() +
    scale_x_dispersion(
      directionality = "conventional",
      n_breaks = 5)
}
```

---

scale_x_fpower *Position scale (x-axis) for Tukey's folded power transformation*

Description

Position scale (x-axis) for Tukey's folded power transformation

Usage

```
scale_x_fpower(lambda = 0, breaks = NULL, labels = NULL, n_breaks = 6, ...)
```

Arguments

lambda	Numeric value of the applied power transformation
breaks	Numeric values indicating where the tick marks should be placed
labels	Character vector giving the labels that should be drawn at the tick marks
n_breaks	Integer specifying the number of tick marks to draw
...	Other arguments passed on to scale_x_continuous()

Details

This function was written with the help of ChatGPT (version GPT-5.1; OpenAI 2025)

Value

The ggplot2 function scale_(x|y)_continuous() with the appropriate transformation

References

OpenAI. (2025). ChatGPT (GPT-5.1) Large language model. https://chat.openai.com

Examples

```
if (require("ggplot2")) {
  ggplot(
    data = data.frame(
      dispersion = seq(0, 1, .01)),
    aes(x = dispersion)) +
    geom_dotplot() +
    scale_x_fpower(lambda = .5)
}
```

---

scale_y_dispersion *Add annotation to tick marks on y-axis (ggplot2) to clarify the directionality of scaling for dispersion scores*

Description

This function can be used when plotting dispersion scores with ggplot2. It forces the y-axis to extend from 0 to 1 and adds verbal information at the endpoints to clarify the directionality of scaling. For conventional scaling, these are "(uneven) 0" and "(even) 1", for gries scaling, these are "(even) 0" and "(uneven) 1".

Usage

```
scale_y_dispersion(directionality, n_breaks = 5, leading_zero = TRUE, ...)
```

Arguments

`directionality`	Character string indicating the directionality of scaling. Must match the way the dispersion scores were calculated. See details below. Possible values are `"conventional"` and `"gries"`
`n_breaks`	Number of major scale breaks: Integer value specifying the number of tick marks to display (default: `5`)
`leading_zero`	Logical. Whether the tick mark labels should include a leading 0 ("0.50") or not (".50"); default is `TRUE`
`...`	Other arguments passed on to `scale_y_continuous()`

Details

This function modifies the y-axis in a ggplot2 object. It forces the axis to extend from 0 to 1 and adds the labels "(even)" and "(uneven)" at the endpoints of the scale (0 and 1), to make clear which value (0 or 1) denotes a maximally even/dispersed/balanced distribution of subfrequencies across corpus parts. The conventional scaling of dispersion measures (see Juilland & Chang-Rodriguez 1964; Carroll 1970; Rosengren 1971) assigns higher values to more even/dispersed/balanced distributions of subfrequencies across corpus parts. In the `{tlda}` package, this is the default setting for all measures. Gries (2008) uses the reverse scaling, with higher values denoting a more uneven/bursty/concentrated distribution; use `directionality = "gries"` to choose this option. The function implements no default, so the user must specify which directionality was used when calculating the scores.

Value

The ggplot2 function scale_y_continuous with the appropriate settings for the argument `limits`, `breaks`, and `labels`.

Author(s)

Lukas Soenning

References

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65. doi:10.1002/j.23338504.1970.tb00778.x

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter. doi:10.1515/9783112415467

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)* 1. 103–127.

Examples

```
if (require("ggplot2")) {
  ggplot(
    data = data.frame(
      frequency = c(1, 1.5, 2, 2.5, 3, 4, 5),
      dispersion = c(.25, .8, .34, .53, .88, .57, .9)),
    aes(x = frequency,
        y = dispersion)) +
    geom_point() +
    scale_y_dispersion(
      directionality = "conventional",
      n_breaks = 3)
}
```

`scale_y_fpower` *Position scale (y-axis) for Tukey's folded power transformation*

Description

Position scale (y-axis) for Tukey's folded power transformation

Usage

```
scale_y_fpower(lambda = 0, breaks = NULL, labels = NULL, n_breaks = 6, ...)
```

Arguments

`lambda`	Numeric value of the applied power transformation
`breaks`	Numeric values indicating where the tick marks should be placed
`labels`	Character vector giving the labels that should be drawn at the tick marks
`n_breaks`	Integer specifying the number of tick marks to draw
`...`	Other arguments passed on to `scale_y_continuous()`

Details

This function was written with the help of ChatGPT (version GPT-5.1; OpenAI 2025)

Value

The ggplot2 function `scale_(x|y)_continuous()` with the appropriate transformation

References

OpenAI. (2025). ChatGPT (GPT-5.1) Large language model. https://chat.openai.com

Examples

```
if (require("ggplot2")) {
  ggplot(
    data = data.frame(
      frequency = rnorm(101),
      dispersion = seq(0, 1, .01)),
    aes(x = frequency,
        y = dispersion)) +
    geom_point() +
    scale_y_fpower(lambda = .5)
}
```

Index