Package 'TTEdata'

January 24, 2020

Title Data sets for TTE course

Version 1.0

Description Data sets for the final project for the class time-to-event analysis of linguistic data.

License do not distribute data

Encoding UTF-8

LazyData yes

RoxygenNote 6.1.1

R topics documented:

`ald` *Auditory lexical decision data*

Description

Auditory lexical decision data from the MALD database (Tucker et al, 2019)

Usage

```
ald
```

Format

A matrix with 22,374 rows and 12 columns:

`word` the item in the auditory lexical decision task

`rt` the average response time in ms

`duration` the acoustic duration of the word, as presented to the participants

`log.frequency` the (log-transformed) frequency of the word in the SUBTLEX-US corpus

`length` the length of the word in letters

`num.phonemes` the length of the word in phonemes

`num.syllables` the length of the word in syllables

`log.old` the (log-transformed) orthographic Levenshtein distance between the word and its 20 closest orthographic neighbors

`log.pld` the (log-transformed) phonological Levenshtein distance between the word and its 20 closest phonological neighbors

`snd` the average semantic similarity between the word and its 5 closest semantic neighbors

`pos` the dominant parts-of-speech category for the word

`sqrt.up` the (square root transformed) uniqueness point of the word; this is the phoneme at which a word a uniquely distinguishable from all other words

Source

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. Behavior Research Methods.

References

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. Behavior Research Methods, 44(4), 991-997.

`hist.english` *Lexical extinction data (English)*

Description

Lexical extinction data for English based on the Google books n-gram data (Michel et al., 2011)

Usage

```
hist.english
```

Format

A matrix with 49,929 rows and 8 columns:

`word` the item in the word naming task

`time` time of the last observation; this equal 2000 if the word remained in the language and the decade in which the word disappeared from the language otherwise

`status` status of the word; 0 if the word remained in the language, 1 if the word disappeared from the language

`log.frequency` the (log-transformed) frequency of the word in the Google n-gram data for the decade from 1800 to 1810

`sqrt.length` the (square root transformed) length of the word in letters

`log.old` the (log-transformed) orthographic Levenshtein distance between the word and its 5 closest orthographic neighbors in 1810

`snd` the average semantic similarity between the word and its 5 closest semantic neighbors in 1810

`pos` the consistency of the mapping from orhography to phonology in 1810

Source

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. science, 331(6014), 176-182.

`hist.german` *Lexical extinction data (German)*

Description

Lexical extinction data for German based on the Google books n-gram data (Michel et al., 2011)

Usage

```
hist.german
```

Format

A matrix with 24,685 rows and 8 columns:

`word` the item in the word naming task

`time` time of the last observation; this equal 2000 if the word remained in the language and the decade in which the word disappeared from the language otherwise

`status` status of the word; 0 if the word remained in the language, 1 if the word disappeared from the language

`log.frequency` the (log-transformed) frequency of the word in the Google n-gram data for the decade from 1800 to 1810

`sqrt.length` the (square root transformed) length of the word in letters

`log.old` the (log-transformed) orthographic Levenshtein distance between the word and its 5 closest orthographic neighbors in 1810

`snd` the average semantic similarity between the word and its 5 closest semantic neighbors in 1810

`pos` the consistency of the mapping from orhography to phonology in 1810

Source

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. science, 331(6014), 176-182.

`hist.russian` *Lexical extinction data (Russian)*

Description

Lexical extinction data for Russian based on the Google books n-gram data (Michel et al., 2011)

Usage

```
hist.russian
```

Format

A matrix with 50,072 rows and 8 columns:

`word` the item in the word naming task

`time` time of the last observation; this equal 2000 if the word remained in the language and the decade in which the word disappeared from the language otherwise

`status` status of the word; 0 if the word remained in the language, 1 if the word disappeared from the language

`log.frequency` the (log-transformed) frequency of the word in the Google n-gram data for the decade from 1800 to 1810

`sqrt.length` the (square root transformed) length of the word in letters

`log.old` the (log-transformed) orthographic Levenshtein distance between the word and its 5 closest orthographic neighbors in 1810

`snd` the average semantic similarity between the word and its 5 closest semantic neighbors in 1810

`pos` the consistency of the mapping from orhography to phonology in 1810

Source

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. science, 331(6014), 176-182.

`ld` *Lexical decision data (aging)*

Description

Lexical decision data for old and young participants from Spieler and Balota (1997)

Usage

```
ld
```

Format

A matrix with 4,422 rows and 8 columns:

- `word` the item in the lexical decision task
- `rt` the average response time in ms
- `age` the age of the participants
- `log.frequency` the (log-transformed) frequency of the word in the SUBTLEX-US corpus
- `length` the length of the word in letters
- `log.old` the (log-transformed) orthographic Levenshtein distance between the word and its 20 closest orthographic neighbors
- `snd` the average semantic similarity between the word and its 5 closest semantic neighbors
- `pos` the dominant parts-of-speech category for the word

Source

Spieler D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. Psychological Science, 8(6), 411-416.

References

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. Behavior Research Methods, 44(4), 991-997.

`ld.chin` *Chinese lexical decision data*

Description

Lexical decision data for Mandarin Chinese from Tsang et al. (2018)

Usage

```
ld.chin
```

Format

A matrix with 9,602 rows and 7 columns:

`word` the item in the lexical decision task

`rt` the average response time in ms

`log.frequency` the (log-transformed) frequency of the word in the CLD

`length` the length of the word in characters

`sqrt.strokes` the (square root transformed) of the number of strokes in the word as a whole

`log.nwf` the (log-transformed) average of the number of words the characters in the word appear in

`snd` the average semantic similarity between the word and its 5 closest semantic neighbors

Source

Tsang, Y. K., Huang, J., Lui, M., Xue, M., Chan, Y. W. F., Wang, S., & Chen, H. C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. Behavior research methods, 50(5), 1763-1777.

References

Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (CLD). Behavior research methods, 50(6), 2606-2629.

`nam` *Word naming data (aging)*

Description

Word naming data for old and young participants from Spieler and Balota (1997)

Usage

```
nam
```

Format

A matrix with 4,422 rows and 8 columns:

`word` the item in the word naming task

`rt` the average response time in ms

`age` the age of the participants

`log.frequency` the (log-transformed) frequency of the word in the SUBTLEX-US corpus

`length` the length of the word in letters

`log.old` the (log-transformed) orthographic Levenshtein distance between the word and its 20 closest orthographic neighbors

`snd` the average semantic similarity between the word and its 5 closest semantic neighbors

`pos` the dominant parts-of-speech category for the word

Source

Spieler D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. Psychological Science, 8(6), 411-416.

References

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. Behavior Research Methods, 44(4), 991-997.

`sd` *Semantic decision data*

Description

Semantic decision data (concrete/abstract) from the Calgary semantic decision project

Usage

```
sd
```

Format

A matrix with 4,422 rows and 8 columns:

- `word` the item in the semantic decision task
- `rt` the average response time in ms
- `log.frequency` the (log-transformed) frequency of the word in the SUBTLEX-US corpus
- `length` the length of the word in letters
- `log.old` the (log-transformed) orthographic Levenshtein distance between the word and its 20 closest orthographic neighbors
- `snd` the average semantic similarity between the word and its 5 closest semantic neighbors
- `pos` the dominant parts-of-speech category for the word
- `type` the semantic type of the word; concrete or abstract
- `concrete.rating` the concreteness rating of the word

Source

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. Behavior Research Methods, 49(2), 407-417.

References

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. Behavior Research Methods, 44(4), 991-997.

vb *Paste tense generation*

Description

Past tense generation data from Cohen et al. (2013)

Usage

```
vb
```

Format

A matrix with 1,978 rows and 7 columns:

`word` the item in the paste tense generation task

`rt` the average response time in ms

`rt` regularity of the verb

`log.frequency` the (log-transformed) frequency of the word in the SUBTLEX-US corpus

`length` the length of the word in letters

`log.old` the (log-transformed) orthographic Levenshtein distance between the word and its 20 closest orthographic neighbors

`snd` the average semantic similarity between the word and its 5 closest semantic neighbors

`pos` the dominant parts-of-speech category for the word

`type` the semantic type of the word; concrete or abstract

`concrete.rating` the concreteness rating of the word

Source

Cohen-Shikora, E. R., Balota, D. A., Kapuria, A., & Yap, M. J. (2013). The past tense inflection project (PTIP): Speeded past tense inflections, imageability ratings, and past tense consistency measures for 2,200 verbs. Behavior research methods, 45(1), 151-159.

References

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. Behavior Research Methods, 44(4), 991-997.

Index