Linguistic Statistics Overview

- **text_id =** text identifier, e.g. 01_03 = the first text for the '0–3 years' age group.

- **num_tokens =** number of tokens ("words"), excluding punctuation.

- **num_sentences =** number of sentences (*sentence* is defined as a sequence of words delimited by punctuation marks).

- **num_types =** number of unique word forms, i.e. the size of the set of tokens considered as distinct types (each word form is counted only once, regardless of how many times it appears). This measure is functional to the calculation of the Type–Token Ratio (TTR).

- **TTR =** *type-token ratio.* It is a simple measure of **lexical diversity**. Higher TTR reflects a higher variety of words, it indicates richer vocabulary.

$$\boldsymbol{TTR = types \div tokens \times 100}$$

- **TTR_lemmas =** *lemma-based type–token ratio.* It is the same measure of lexical diversity, but calculated on lemmas rather than surface word forms. A higher TTR_lemmas reflects a richer vocabulary after reducing inflectional variation, since different word forms of the same lemma are counted as one. (*la, gli, i, le, etc.* → *"il"*).

- **avg_sentences_length =** average sentence length in tokens, i.e. the mean number of tokens per sentence in a text.

- **zipf_mean =** average Zipf frequency of the words in a text. The Zipf scale measures word frequency on a logarithmic scale ranging from 1 (very low frequency) to 7 (very high frequency).

Zipf value	fpmw	Examples
1	0.01	antifungal, bioengineering, farsighted, harelip, proofread
2	0.1	airstream, doorkeeper, neckwear, outsized, sunshade
3	1	beanstalk, cornerstone, dumpling, insatiable, perpetrator

Zipf value	fpmw	Examples
4	10	dirt, fantasy, muffin, offensive, transition, widespread
5	100	basically, bedroom, drive, issues, period, spot, worse
6	1000	day, great, other, should, something, work, years
7	10,000	and, for, have, I, on, the, this, that, you

(*fpmw* = frequency per million words)

Taken from Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: *A New and Improved Word Frequency Database for British English. Quarterly Journal of Experimental Psychology*, 67(6), 1176-1190. https://doi.org/10.1080/17470218.2013.850521 (Original work published 2014)

- **percent_rare_words =** percentage of rare words in a text. Rare words are defined as those with a Zipf frequency value lower than 3.

- **Gulpease_index =** index of readability. It's a formula specifically designed for Italian texts. It produces a score between 0 (*very difficult*) and 100 (*very easy*), based on sentence length and word length in letters.

$$\textbf{Gulpease Index} = 89 + \frac{300 \times \textbf{number of sentences} - 10 \times \textbf{number of letters}}{\textbf{number of words}}$$

(https://it.wikipedia.org/wiki/Indice_Gulpease)

- **top_lemmas =** the 20 most frequent lemmas in a text, ranked by frequency.

- **top_bigram =** the 10 most frequent bigrams (*two-word sequences*) in a text, ranked by frequency.

All the following measures are reported as **percentages** relative to the total number of tokens in the text, in order to normalize across texts of different lengths.

- **num_function_words =** number of function words in a text. *Function words* are defined as tokens whose POS tag belongs to the set:
  - ADP *adposition*
  - AUX *auxiliary verb*

  - CCONJ *coordinating conjunction*
  - SCONJ *subordinating conjunction*
  - DET *determiner*
  - PRON *pronoun*
  - PART *particle*
  - INTJ *interjection*

- **num_content_words =** number of content words in a text. *Content words* are defined as tokens whose POS tag belongs to the set:
  - NOUN *noun*
  - VERB *verb*
  - ADJ *adjective*
  - ADV *adverb*
  - PROPN *proper noun*

- **num_verbs =** number of verbs in a text.

- **num_adjectives =** number of adjectives in a text.

- **num_nouns =** number of nouns in a text.