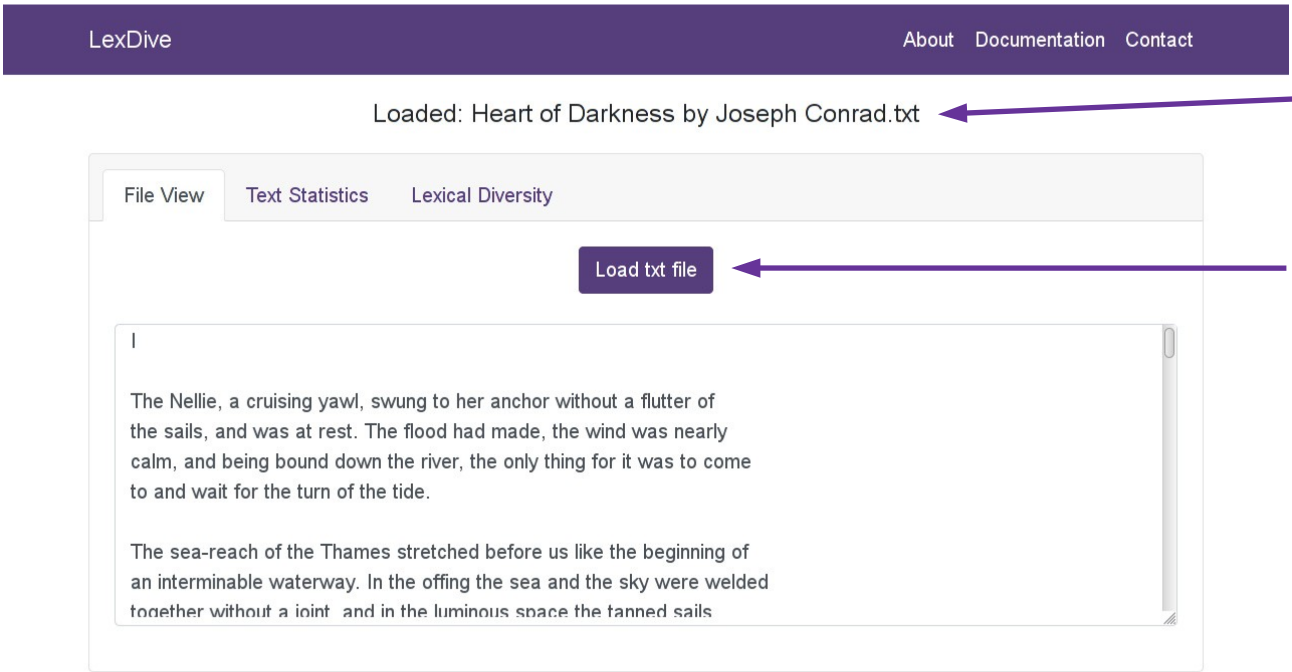
LexDive, version 1.3
A program for counting lexical diversity
Developed by Łukasz Stolarski, December 2020
email: lukasz.stolarski@ujk.edu.pl

The interface of the program includes three tabs:

1 File View tab – you should start working with LexDive by loading a plain text file. You can do this either by choosing "File>Load txt file" in the menu, or pressing the "Load txt file" as show in the picture below. The text of the file will be displayed in the window at the bottom. To use formats other than "txt", choose one of the conversion tools available on the Internet or simply copy the text you want to process into a text editor and save it as a "txt" file.



Loaded: Heart of Darkness by Joseph Conrad.txt

I

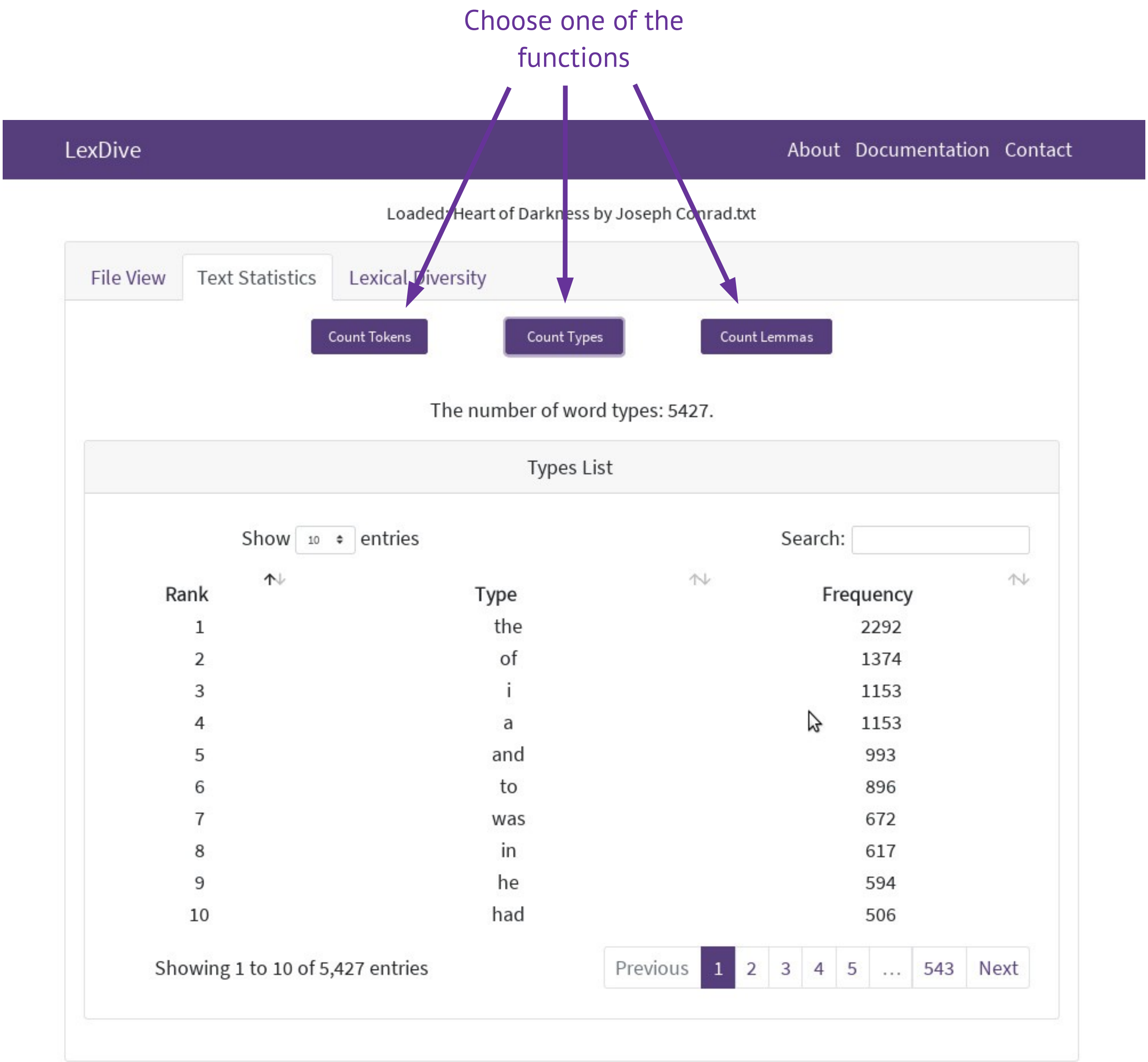
The Nellie, a cruising yawl, swung to her anchor without a flutter of
the sails, and was at rest. The flood had made, the wind was nearly
calm, and being bound down the river, the only thing for it was to come
to and wait for the turn of the tide.

The sea-reach of the Thames stretched before us like the beginning of
an interminable waterway. In the offing the sea and the sky were welded
together without a joint, and in the luminous space the tanned sails

the name of the file which has been loaded

press this button to load a txt file or paste text into the windo below

2 Text Statistics tab – the three options available in this tab allow the user to examine the basic statistics of the text. The "Count tokens" function calculates all the word tokens in the entire text. The "Count types" option calculates the number of word types and the results are arranged according to the frequency of each type. Additionally, the "Count lemmas" function displays the results arranged according to the frequency of lemmas obtained (see the picture below).



Choose one of the functions

Loaded: Heart of Darkness by Joseph Conrad.txt

The number of word types: 5427.

Types List

| Rank | Type | Frequency |
|---|---|---|
| 1 | the | 2292 |
| 2 | of | 1374 |
| 3 | i | 1153 |
| 4 | a | 1153 |
| 5 | and | 993 |
| 6 | to | 896 |
| 7 | was | 672 |
| 8 | in | 617 |
| 9 | he | 594 |
| 10 | had | 506 |

Showing 1 to 10 of 5,427 entries

3 Lexical Diversity tab – before the program can calculate lexical diversity, the user needs to choose a lexical diversity index. If the index is other than MTLD or HDD, a sampling method can be changed from "whole text" to either "split text" or "equal text" (for

3.1 Lexical Diversity indices – at the current stage of development, the program uses the following indices:

3.1.1 MTLD (Measure of Textual Lexical Diversity) – the most reliable index of lexical diversity available, highly robust to differences in text length. The method may be used only with "Whole text" sampling because it executes its own segmentation which is dependent on the lexical diversity of the text itself (for details see McCarthy & Jarvis, 2010).

3.1.2 MTLD (lemmas) – the version of MTLD in which lemmas are used instead of types. It may be more precise in estimating actual lexical diversity than the basic version of the method.

3.1.3 HDD (Hypergeometric distribution D) – this lexical diversity index is provided in Python “lexical-diversity” package[1]. According to the author of the library, it is “a more straightforward and reliable implementation of vocD (Malvern, Richards, Chipere, & Durán, 2004) as per McCarthy & Jarvis (2007, 2010)”.

3.1.4 TTR (Text Token Ratio) - it is the most popular index of lexical diversity. It is calculated by dividing the number of word types by the number of word tokens. It is heavily affected by text length differences, so it should not be used with “Whole text” sampling when comparing texts of different lengths. Instead, “Equal text” or, preferably, “Split text” should be applied.

$$TTR = \frac{Types}{Tokens}$$

3.1.5 TTR (lemmas) - the version of TTR in which lemmas are used instead of types. It may be more precise in estimating actual lexical diversity than the basic version of the method.

3.1.6 Algebraic transformations of TTR (as reported in Jarvis (2002))

3.1.6.1 Herdan’s C

$$C = \frac{\log Types}{\log Tokens}$$

3.1.6.2 Herdan’s C (lemmas)- the version of Herdan’s C in which lemmas are used instead of types. It may be more precise in estimating actual lexical diversity than the basic version of the method.

3.1.6.3 Guiraud’s R

$$R = \frac{Types}{\sqrt{Tokens}}$$

3.1.6.4 Guiraud’s R (lemmas)- the version of Guiraud’s R in which lemmas are used instead of types. It may be more precise in estimating actual lexical diversity than the basic version of the method.

3.1.6.5 Uber U

[1] https://github.com/kristopherkyle/lexical_diversity

$$U = \frac{(logTokens)^2}{(logTokens - logTypes)}$$

3.1.6.6 Uber U (lemmas) - the version of Uber U in which lemmas are used instead of types. It may be more precise in estimating actual lexical diversity than the basic version of the method.

3.2 Sampling methods

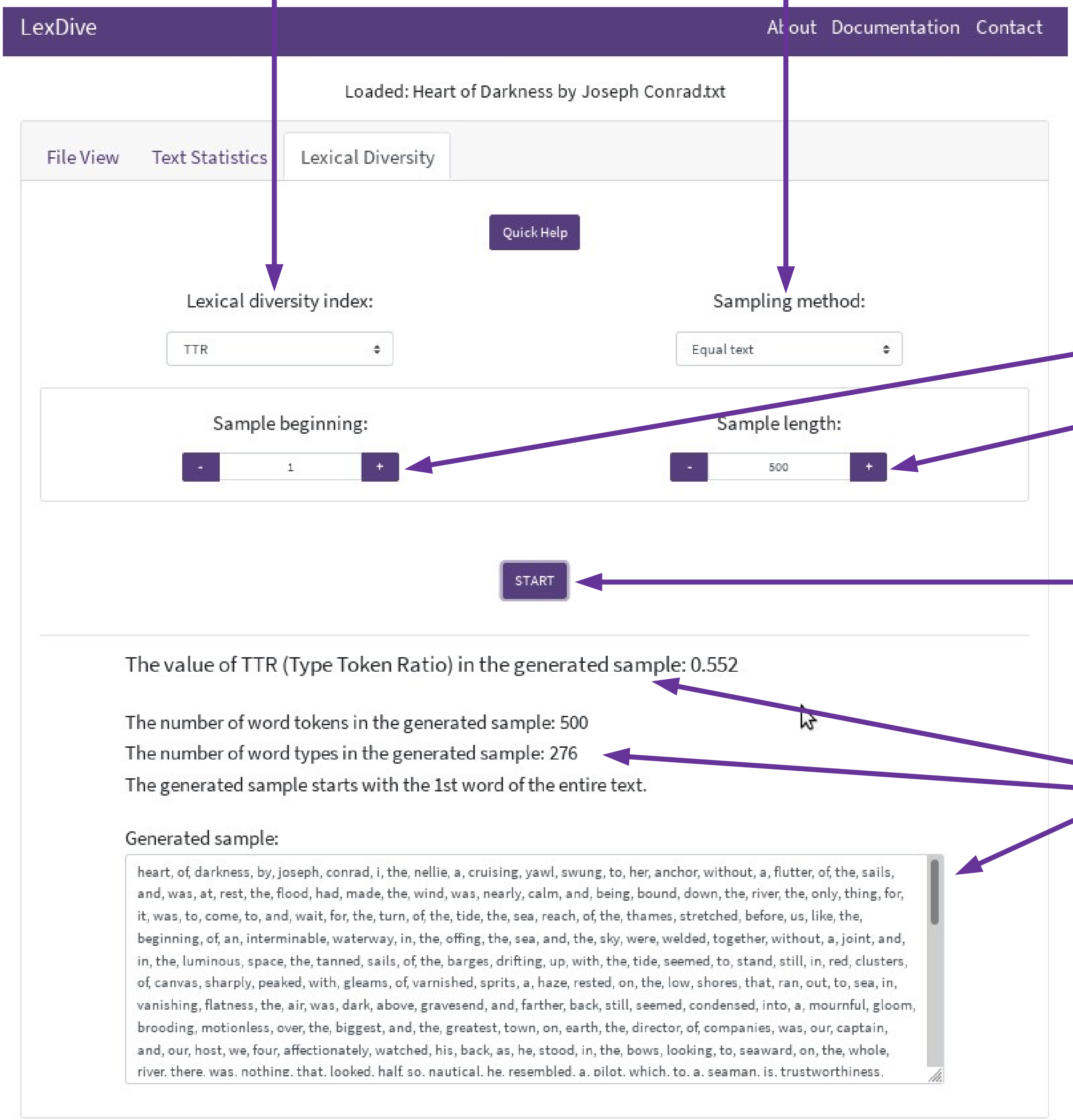
3.2.1 “Whole text” - the program allows the user to calculate lexical diversity in the entire text (this is the only available sampling method for any version of MTLD and HDD). When comparing two of more texts of different lengths, some lexical diversity indices, such as TTR, should not be used in combination with this method, because their values decrease as the size of the text increases, regardless of the overall author’s style of writing (cf. Broeder, Extra, & Van Hout, 1986; Johansson, 1999, 2009).

3.2.2 “Split text” – not available for any version of MTLD and HDD. A sampling method which makes it possible to compare lexical diversity of different texts when using indices which are strongly affected by text length (TTR and all its alternations listed on pages 3-4). It involves taking random samples from the text. LexDive allows the user to choose both the size of the intended text size and also the size of subsamples. The program calculates the optimal gaps between each subsample according to the equation *(nt–ns)/ss*, where *nt* = the number of tokens in the whole text, *ns* = the intended number of tokens in the sample, and *ss* = the number of sub-samples (which is calculated from *ns/sss*, where *ns* = the intended number of tokens in the sample, and *sss* = the intended size of the sub-samples). When comparing texts of different lengths, the samples should be of the same length. The size of subsamples, on the other hand, does not influence the results in a significant way, although it should not be very small, especially if the difference between the sample size and the whole text size is not large. For instance, if we want a sample of 30 000 from a text containing 40 000 word tokens, we may take sub-samples which contain 15 word tokens each (the optimal gap between the sub-samples will be 5 word tokens in the original text). However, a sample of 39 000 word tokens taken from the same text would require much larger sub-samples (the gap between the sub-samples containing 15 words would be 0, so the last 1000 word tokens in the original text would not be properly represented in the sample).

3.2.3 “Equal text” - not available for any version of MTLD and HDD. Another sampling method which makes it possible to compare lexical diversity of different texts when using indices which are strongly affected by text length (TTR and all its alternations listed on pages 3-4). It requires specifying the starting point of the sample and the size of the sample.

Choose a lexical diversity index

Choose a sampling method (if the index chosen is any version of MTLD or HDD, only “Whole text” is available)



Loaded: Heart of Darkness by Joseph Conrad.txt

Lexical diversity index: TTR

Sampling method: Equal text

Sample beginning: 1

Sample length: 500

The value of TTR (Type Token Ratio) in the generated sample: 0.552

The number of word tokens in the generated sample: 500
The number of word types in the generated sample: 276
The generated sample starts with the 1st word of the entire text.

Generated sample:

heart, of, darkness, by, joseph, conrad, i, the, nellie, a, cruising, yawl, swung, to, her, anchor, without, a, flutter, of, the, sails, and, was, at, rest, the, flood, had, made, the, wind, was, nearly, calm, and, being, bound, down, the, river, the, only, thing, for, it, was, to, come, to, and, wait, for, the, turn, of, the, tide, the, sea, reach, of, the, thames, stretched, before, us, like, the, beginning, of, an, interminable, waterway, in, the, offing, the, sea, and, the, sky, were, welded, together, without, a, joint, and, in, the, luminous, space, the, tanned, sails, of, the, barges, drifting, up, with, the, tide, seemed, to, stand, still, in, red, clusters, of, canvas, sharply, peaked, with, gleams, of, varnished, sprits, a, haze, rested, on, the, low, shores, that, ran, out, to, sea, in, vanishing, flatness, the, air, was, dark, above, gravesend, and, farther, back, still, seemed, condensed, into, a, mournful, gloom, brooding, motionless, over, the, biggest, and, the, greatest, town, on, earth, the, director, of, companies, was, our, captain, and, our, host, we, four, affectionately, watched, his, back, as, he, stood, in, the, bows, looking, to, seaward, on, the, whole, river, there, was, nothing, that, looked, half, so, nautical, he, resembled, a, pilot, which, to, a, seaman, is, trustworthiness,

If the sampling method is "Split text" or “Equal text”, specify additional parameters

Press START to begin processing

The major results and additional information are displayed here

## 4 Legal matter

LexDive can be used freely for non-profit purposes. For commercial use, contact the author.
The software comes on an 'as is' basis, and the author will accept no liability for any damage that may result from using the software.

## References:

Broeder, P., Extra, G., & Van Hout, T. (1986). Measuring lexical richness and diversity in second language research. *Polyglot*, *8*, 1–16.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84.

Johansson, V. (1999). Word frequencies in speech and writing: a study of expository discourse. In *Working papers in developing literacy across genres, modalities, and languages* (Vol. I, pp. 182–198). Tel Aviv: Tel Aviv University Press.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics*, *53*, 61–79.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). Lexical diversity and language development. *Houndmills, Hampshire, UK: Palgrave Macmillan*.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.