




TextDescriptives: A Python package for calculating a large variety of metrics from text

Lasse Hansen ^{1,2,3¶}, Ludvig Renbo Olsen ^{4,2}, and Kenneth Enevoldsen ^{2,3}

¹ Department of Affective Disorders, Aarhus University Hospital - Psychiatry, Aarhus, Denmark ² Department of Clinical Medicine, Aarhus University, Aarhus, Denmark ³ Center for Humanities Computing, Aarhus University, Aarhus, Denmark ⁴ Department of Molecular Medicine (MOMO), Aarhus University, Aarhus, Denmark ¶ Corresponding author

DOI: [10.21105/joss.05153](https://doi.org/10.21105/joss.05153)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: Fabian Scheipl 

Reviewers:

- [@RichardLitt](#)
- [@linuxscout](#)

Submitted: 06 January 2023

Published: 24 April 2023

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

Natural language processing (NLP) tasks often require a thorough understanding and description of the corpus. Document-level metrics can be used to identify low-quality data, assess outliers, or understand differences between groups. Further, text metrics have long been used in fields such as the digital humanities where e.g. metrics of text complexity are commonly used to analyse, understand and compare text corpora. However, extracting complex metrics can be an error-prone process and is rarely rigorously tested in research implementations. This can lead to subtle differences between implementations and reduces the reproducibility of scientific results.

TextDescriptives offers a simple and modular approach to extracting both simple and complex metrics from text. It achieves this by building on the spaCy framework ([Honribal et al., 2020](#)). This means that TextDescriptives can easily be integrated into existing workflows while leveraging the efficiency and robustness of the spaCy library. The package has already been used for analysing the linguistic stability of clinical texts ([Hansen et al., 2022](#)), creating features for predicting neuropsychiatric conditions ([Hansen et al., 2023](#)), and analysing linguistic goals of primary school students ([Tannert, 2023](#)).

Statement of need

Computational text analysis is a broad term that refers to the process of analyzing and understanding text data. This often involves calculating a set of metrics that describe relevant properties of the data. Dependent on the task at hand, this can range from simple descriptive statistics related to e.g. word or sentence length to complex measures of text complexity, coherence, or quality. This often requires drawing on multiple libraries and frameworks or writing custom code. This can be time-consuming and prone to bugs, especially with more complex metrics.

TextDescriptives seeks to unify the extraction of document-level metrics, in a modular fashion. The integration with spaCy allows the user to seamlessly integrate TextDescriptives in existing pipelines as well as giving the TextDescriptives package access to model-based metrics such as dependency graphs and part-of-speech tags. The ease of use and the variety of available metrics allows researchers and practitioners to extend the granularity of their analyses within a tested and validated framework.

Implementations of the majority of the metrics included in TextDescriptives exist, but none as feature complete. The textstat library ([Ward, 2022](#)) implements the same readability metrics, however, each metric has to be extracted one at a time with no interface for multiple extractions. spacy-readability ([Holtzsch, 2019](#)) adds readability metrics to spaCy pipelines, but does

not work for new versions of spaCy ($\geq 3.0.0$). The textacy (DeWilde, 2021) package has some overlap with TextDescriptives, but with a different focus. TextDescriptives focuses on document-level metrics, and includes a large number of metrics not included in textacy (dependency distance, coherence, and quality), whereas textacy includes components for preprocessing, information extraction, and visualization that are outside the scope of TextDescriptives. What sets TextDescriptives apart is the easy access to document-level metrics through a simple user-facing API and exhaustive documentation.

Features & Functionality

TextDescriptives is a Python package and provides the following spaCy pipeline components: `textdescriptives.descriptive_stats`: Calculates the total number of tokens, number of unique tokens, number of characters, and the proportion of unique tokens, as well as the mean, median, and standard deviation of token length, sentence length, and the number of syllables per token. `textdescriptives.readability`: Calculates the Gunning-Fog index, the SMOG index, Flesch reading ease, Flesch-Kincaid grade, the Automated Readability Index, the Coleman-Liau index, the Lix score, and the Rix score. `textdescriptives.dependency_distance`: Calculates the mean and standard deviation of the dependency distance (the average distance between a word and its head word), and the mean and the standard deviation of the proportion adjacent dependency relations on the sentence level. `textdescriptives.pos_proportions`: Calculates the proportions of all part-of-speech tags in the documents. `textdescriptives.coherence`: Calculates the first- and second-order coherence of the document based on word embedding similarity between sentences. `textdescriptives.quality`: Calculates the text-quality metrics proposed in Rae et al. (2022) and Raffel et al. (2020). These measures can be used for filtering out low-quality text prior to model training or text analysis. These include heuristics such as the number of stop words, ratio of words containing alphabetic characters, proportion of lines ending with an ellipsis, proportion of lines starting with a bullet point, ratio of symbols to words, and whether the document contains a specified string (e.g. “lorem ipsum”), as well as repetitious text metrics such as the proportion of lines that are duplicates, the proportion of paragraphs in a document that are duplicates, the proportion of n-gram duplicates, and the proportion of characters in a document that are contained within the top n-grams.

All the components can be added to an existing spaCy pipeline with a single line of code, and jointly extracted to a dataframe or dictionary with a single call to `textdescriptives.extract_{df|dict}(doc)`.

Example Use Cases

Descriptive statistics can be used to summarize and understand data, such as by exploring patterns and relationships within the data, getting a better understanding of the data set, or identifying any changes in the distribution of the data. Readability metrics, which assess the clarity and ease of understanding of written text, have a variety of applications, including the design of educational materials and the improvement of legal or technical documents (DuBay, 2004). Dependency distance can be used as a measure of language comprehension difficulty or of sentence complexity and has been used for analysing properties of natural language or for similar purposes as readability metrics (Gibson et al., 2019; Liu, 2008). The proportions of different parts of speech in a document have been found to be predictive of certain mental disorders and can also be used to assess the quality and complexity of text (Tang et al., 2021). Semantic coherence, or the logical connection between sentences, has primarily been used in the field of computational psychiatry to predict the onset of psychosis or schizophrenia (Bedi et al., 2015; Parola et al., 2022), but it also has other applications in the digital humanities. Measures of text quality are useful cleaning and identifying low-quality data (Rae et al., 2022; Raffel et al., 2020).

Target Audience

The package is mainly targeted at NLP researchers and practitioners. In particular, researchers from fields new to NLP such as the digital humanities and social sciences as researchers might benefit from the readability metrics as well as the more complex, but highly useful, metrics such as coherence and dependency distance.

Acknowledgements

The authors thank the [contributors](#) of the package including Martin Bernstorff for his work on the part-of-speech component, and Frida Hæstrup and Roberta Rocca for important fixes. The authors would also like to Dan Sattrup Nielsen for helpful reviews on early iterations of the text quality implementations.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophrenia*, 1(1), 1–7. <https://doi.org/10.1038/npjschz.2015.30>

DeWilde, B. (2021). *Textacy: NLP, before and after spaCy* (Version 0.12.0). <https://github.com/chartbeat-labs/textacy>

DuBay, W. H. (2004). The principles of readability. *Online Submission*.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5), 389–407.

Hansen, L., Enevoldsen, K., Bernstorff, M., Perfalk, E., Danielsen, A. A., Nielbo, K. L., & Østergaard, S. D. (2022). Lexical stability of psychiatric clinical notes from electronic health records over a decade. *medRxiv*, 2022.09.05.22279610. <https://doi.org/10.1101/2022.09.05.22279610>

Hansen, L., Rocca, R., Simonsen, A., Parola, A., Bliksted, V., Ladegaard, N., Bang, D., Tylén, K., Weed, E., Østergaard, S. D., & Fusaroli, R. (2023). *Automated speech- and text-based classification of neuropsychiatric conditions in a multidagnostic setting*. *arXiv:2301.06916*. <https://doi.org/10.48550/arXiv.2301.06916>

Holtzsch, M. (2019). *Spacy-readability: spaCy pipeline component for adding text readability meta data to doc objects*. (Version 1.4.1).

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in python*. <https://doi.org/10.5281/zenodo.1212303>

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), 159–191. <https://doi.org/10.17791/jcs.2008.9.2.159>

Parola, A., Lin, J. M., Simonsen, A., Bliksted, V., Zhou, Y., Wang, H., Inoue, L., Koelkebeck, K., & Fusaroli, R. (2022). Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of NLP automated measures of coherence. *Schizophrenia Research*. <https://doi.org/10.1016/j.schres.2022.07.002>

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G. van den, Hendricks, L. A., Rauh, M., Huang, P.-S., ... Irving, G. (2022). *Scaling language models: Methods, analysis & insights from training gopher* (No. arXiv:2112.11446). *arXiv*. <https://doi.org/10.48550/arXiv.2112.11446>

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

- arXiv:1910.10683 [Cs, Stat]*. <http://arxiv.org/abs/1910.10683>
- Tang, S. X., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R. E., Bhati, M. T., Wolf, D. H., Sedoc, J., & Liberman, M. Y. (2021). Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *Npj Schizophrenia*, 7(1), 1–8. <https://doi.org/10.1038/s41537-021-00154-3>
- Tannert, M. (2023). *Skriftsproglig udvikling i grundskolens danskfag* [PhD thesis]. Aarhus University.
- Ward, A. (2022). *Textstat*. Textstat. <https://github.com/textstat/textstat>