

pyliwc: Linguistic Inquiry and Word Count in Python

Camille Lacan ¹

¹ IAE School of Management, University Perpignan Via Domitia, FR

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

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

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright
and release the work under a
Creative Commons Attribution 4.0
International License ([CC BY 4.0](#)).

Summary

Linguistic Inquiry and Word Count (LIWC), developed by Pennebaker et al. (2015), is a widely recognized tool for analyzing word usage through a dictionary-based approach. LIWC provides insights by counting word occurrences in texts and outputs the percentage of words that fall into one or more of over 80 linguistic (e.g., first-person singular pronouns, conjunctions), psychological (e.g., anger, achievement), and topical (e.g., leisure, money) categories. Researchers extensively use LIWC in social science research and linguistic studies to examine language. However, LIWC's reliance on third-party proprietary software makes it challenging to analyze large datasets such as social media conversations efficiently.

pyliwc addresses this limitation by allowing the use of the LIWC dictionary in Python. pyliwc includes functions to run LIWC analysis with internal or custom dictionaries, enabling efficient text analysis. This package handles large text corpora from Pandas DataFrames or text files, making it efficient for text analysis. Finally, pyliwc includes additional features like language style matching and narrative arc analysis, making it a comprehensive tool for linguistic and social science research.

Licensing requirements

The “pyliwc” Python package serves as an interface for using the LIWC software and its dictionaries. To utilize this package effectively, users must have a valid academic LIWC license. This license is necessary to access both the proprietary LIWC dictionaries and any custom dictionaries created by users.

Please ensure that you have acquired the appropriate LIWC software license before attempting to use the “pyliwc” package.

Please visit the [LIWC website](#) for installation instructions.

Statement of need

The rise of digital communication has led to an explosion of unstructured data, with text accounting for approximately 80-90% of this data (Boegershausen et al., 2022). This includes data from social media posts, emails, forums, and other online interactions. Analyzing this vast amount of textual data is challenging due to the unstructured nature of such data, which complicates the extraction of meaningful insights (Berger et al., 2022). Traditional approaches to text analysis often struggle to scale with the growing volume and complexity of data (Humphreys & Wang, 2017). A variety of software tools are available for text analysis, particularly in the realm of sentiment analysis.

Existing Python packages include:

- TextBlob: Provides basic sentiment analysis and text processing capabilities.

- VADER (Valence Aware Dictionary and sEntiment Reasoner): Specializes in sentiment analysis, particularly for social media texts.
- spaCy: Offers advanced NLP features, including sentiment analysis through third-party extensions.
- NLTK (Natural Language Toolkit): A comprehensive library for text processing with support for sentiment analysis via custom implementations.

In linguistics, one widely used set of dictionaries is the Linguistic Inquiry and Word Count – LIWC (Boyd et al., 2022). The Linguistic Inquiry and Word Count (LIWC) dictionary is a prominent tool in text analysis. Developed by Pennebaker et al. (2015), LIWC categorizes words into various linguistic, psychological, and topical categories (Boyd et al., 2022). These include linguistic features (measures of verb tense, sentence structure, and other grammatical aspects), psychological categories (emotions, cognition, and social processes), substantive categories (e.g., leisure and money), and analysis of punctuation usage. LIWC's dictionaries have been validated on a range of materials such as academic abstracts, literature texts, and other written corpora.

However, LIWC requires users to interact directly with its graphical user interface (GUI) for data analysis, leading to potential inefficiencies and errors in data handling. Users must export data from Python, process it in LIWC, and then re-import the results, which can be cumbersome and error-prone. Additionally, LIWC's GUI does not easily allow changes to dictionaries or manage large-scale data analysis effectively.

Thus, there is a significant need for researchers and data scientists who require the advanced capabilities of the LIWC dictionary but seek to integrate it directly into their Python workflows (Berger et al., 2019, 2022). `pyliwc` addresses these challenges by integrating LIWC's dictionary-based analysis directly in Python. This package eliminates the need for external software interaction, supports custom dictionary usage, and enhances the efficiency of processing large datasets. By streamlining the analysis process, `pyliwc` offers a practical and scalable solution for researchers and practitioners working with extensive textual data.

Main features

The package offers a wide range of features, including:

- LIWC Text Analysis:
 - Analyze text data from various sources, including CSV files, directories, Pandas DataFrames, and individual strings.
 - Supports internal dictionaries (e.g., LIWC22, LIWC2015) as well as custom dictionaries.
 - Output results directly in a convenient Pandas DataFrame for easy integration with other data processing tools.
- Linguistic Style Matching (LSM):
 - Perform person- and group-level LSM analysis using a DataFrame to evaluate the alignment of linguistic styles in conversational data.
 - Supports pairwise LSM calculations for detailed analysis of interpersonal communication dynamics.
- Narrative Arc Analysis:
 - Analyze the narrative arc of text data to understand staging, progression, and cognitive tension, offering deep insights into storytelling elements.
 - Includes graphical capabilities, allowing users to visualize narrative structures: staging, plot progression, and cognitive tension over time.

- 84 – Provides customizable scaling methods and segment options for precise control
- 85 over the analysis process.
- 86 ▪ Integration with LIWC CLI:
 - 87 – Seamlessly execute LIWC commands and capture output for further processing,
 - 88 leveraging the full power of LIWC’s linguistic analysis capabilities. This feature
 - 89 includes multithreading support for improved performance and faster analysis across
 - 90 large datasets.
- 91 ▪ Output Options:
 - 92 – Flexible output formats, including CSV, JSON, and direct integration with Pandas
 - 93 DataFrames, ensuring compatibility with a wide range of data analysis workflows.

94 **Example of use: Analyzing U.S. Presidential Inaugural Addresses**

95 This short example illustrates how to use the `pyliwc` package to analyze the linguistic styles
96 and psychological attributes of inaugural addresses from four U.S. Presidents: George W. Bush,
97 Barack Obama, Donald Trump, and Joe Biden.

```
from pyliwc import Liwc
import pandas as pd

# Initialize with the LIWC-22 dictionary
liwc = Liwc('LIWC-22-cli')

# Read the CSV file containing U.S. Presidents' speeches
df = pd.read_csv('../data/US-president.csv')

# Analyze the text data using pyliwc
result_df = liwc.analyze_df(df['Text'], liwc_dict='LIWC22')

# Print the results
print(result_df)
```

98 **Narrative Arc Analysis** aims to analyze the structure and flow of narratives within text data
99 (Boyd et al., 2020). It provides insights into how narratives evolve over time and helps to
100 identify the progression of thematic elements within a text.

```
# Analyze the narrative arc of the speeches
arc = liwc.narrative_arc(
    df=df,
    column_names=['Text'],
    output_individual_data_points=0,
    scaling_method='0-100',
    segments_number=5
)

# Results of the narrative arc analysis
print(arc)

# Plot the results of the LIWC analysis
liwc.plot_narrative_arc(df=arc)
```

101 The output of the narrative arc analysis is presented in the Table 1 and illustrated in the Figure
102 1.

Table 1: Results of Narrative Arc Analysis

Table with 6 columns: President, WC, Overall, Staging, Progression, and Cognitive Tension. Rows include Bush, Obama, Trump, and Biden with their respective values.

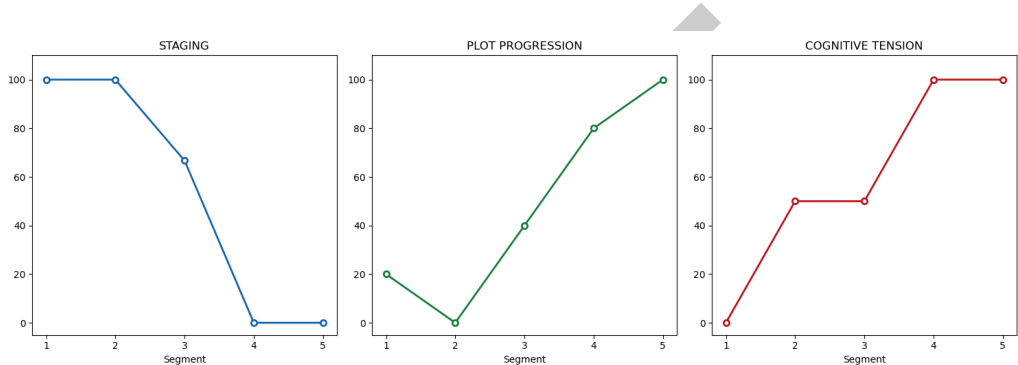


Figure 1: Narrative Arc of U.S.Presidential Inaugural Addresses

Conclusion

The LIWC framework has facilitated numerous research projects across the social sciences, demonstrating its broad applicability (Berger et al., 2019; Humphreys & Wang, 2017). By leveraging LIWC's comprehensive dictionaries, pylwc assists researchers in advancing linguistic research, conducting large-scale analyses, and deepening our understanding of human language.

Availability and Documentation

pylwc is available under the MIT License. The library may be cloned from the GitHub repository, or can be installed by pip: pip install pylwc. Documentation is provided via Quarto, including a quickstart introducing major functionality and a detailed API reference. Extensive unit testing is employed throughout the library.

References

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2019). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1), 1–25. <https://doi.org/10.1177/0022242919873106>

Berger, J., Packard, G., Boghrati, R., Hsu, M., Humphreys, A., Luangrath, A., Moore, S., Nave, G., Olivola, C., & Rocklage, M. (2022). Marketing insights from text analysis. *Marketing Letters*. <https://doi.org/10.1007/s11002-022-09635-6>

Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of gold: Scraping web data for marketing insights. *Journal of Marketing*, 86(5), 1–20. <https://doi.org/10.1177/00222429221100750>

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 10.

- 126 Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core
127 narrative structures through text analysis. *Science Advances*, 6(32). <https://doi.org/10.1126/sciadv.aba2196>
128
- 129 Humphreys, A., & Wang, R. J.-H. (2017). Automated text analysis for consumer research.
130 *Journal of Consumer Research*, 44(6), 1274–1306. <https://doi.org/10.1093/jcr/ucx104>
- 131 Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and*
132 *word count: LIWC 2015* (Pennebaker Conglomerates).

DRAFT