Corpus Linguistics in LuaLaTeX

Statistics, Tables, and Zipf Plots

Your Name

October 31, 2026

This worked example shows how to produce publication-ready statistical output directly in LuaLaTeX: aligned decimal tables with `siunitx`, and vector-based plots with `pgfplots`. No Python screenshots required — everything stays in the `.tex` source and recompiles when the data change.

1 Aligned Statistical Tables with `siunitx`

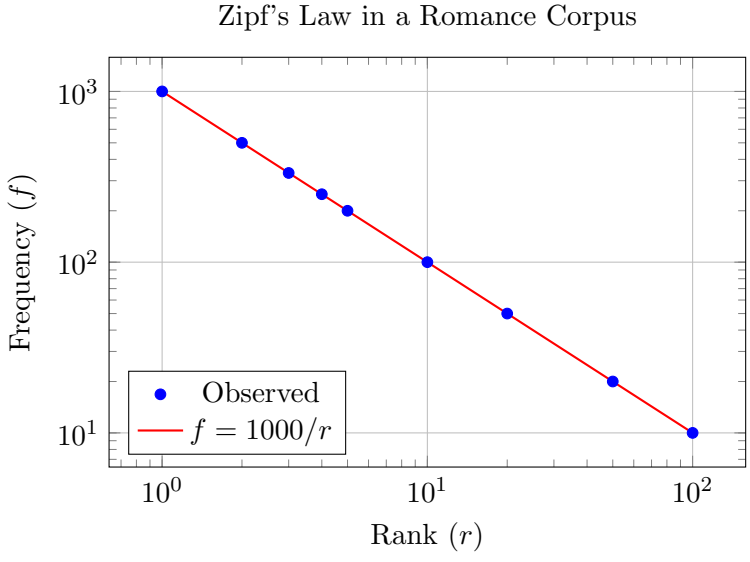
The `S` column type in `siunitx` aligns numbers on their decimal point. Scientific notation is formatted automatically, and the p in p-value stays correctly italicized.

Language	**Mean Word Length**	**p-value**
French	4.812	3.4×10^{-5}
Spanish	5.123	1.2×10^{-6}
Italian	5.401	4.5×10^{-4}

In running text you can also use `\num{1.23e-4}` to format a number as 1.23×10^{-4}, and `\SI{95}{\percent}` for a value with its unit: 95 %.

2 Zipf's Law Plot with `pgfplots`

The plot below is generated purely in LaTeX from inline coordinates. The output is vector-based, so it stays sharp at any zoom level and the fonts match the rest of the document.



The red curve is the theoretical Zipf law $f = C/r$ for comparison. Replace the inline coordinates with your own corpus data when you have them.

3 Loading Data from a CSV

For a real corpus you do not want to retype counts into the TEX source. Load them from a CSV with `pgfplotstable`:

```
\usepackage{pgfplotstable}
% ...
\addplot table [x=rank, y=freq, col sep=comma] {my_corpus.csv};
```

The CSV needs a header row (e.g. `rank,freq`) and one data row per observation. Every recompile re-reads the file, so updating the corpus updates the plot without touching the LaTeX source.

4 Short Reference

- `\num{1.23e-4}` — format a number in text.
- `S[table-format=2.3]` — column type that aligns decimals; `2.3` means up to 2 digits before and 3 after.
- `\SI{95}{\percent}` — number with a unit.
- `\begin{loglogaxis}` — use instead of `axis` for log-log plots.
- `\addplot table [x=..., y=...]` — read columns from a CSV.