Revista Ibero-Americana de Humanidades, Ciências e Educação — REASE

doi.org/10.51891/rease.v12i9.30297

A COMPACT EXTENSION OF HEAPS' LAW: INCORPORATING PHONOLOGICAL STRUCTURE AND LEXICAL REPETITION

UMA EXTENSÃO COMPACTA DA LEI DE HEAPS: INCORPORANDO A ESTRUTURA FONOLÓGICA E A REPETIÇÃO LEXICAL

UNA EXTENSIÓN COMPACTA DE LA LEY DE HEAPS: INCORPORANDO LA ESTRUCTURA FONOLÓGICA Y LA REPETICIÓN LÉXICA

Raquel Romes Linhares[1]
Regis Nunes Vargas[2]

ABSTRACT: The classical formulation of Heaps' Law establishes a bivariate relationship between the length of a text and the growth of its vocabulary, disregarding morphological, phonological, and stylistic factors inherent to natural language. This work proposes and validates a compact, generalized extension of Heaps' Law, incorporating the proportion of monosyllabic words (p_1), theoretically grounded in Zipf's Law of Abbreviation, and a variable indicating high lexical repetition (Drep). The simulation-based validation adopts a mechanistic data-generating process (DGP) based on a Zipf-Mandelbrot urn, in which vocabulary and the proportion of monosyllables emerge from a token sampling process, without imposing the functional form of either compared model on the generating process. The results tables report the average estimated parameters across 200 Monte Carlo replications, ensuring that the estimates do not depend on a single sample realization. Empirical validation on a large English-language dataset, the Movie Reviews Corpus, corroborates the gains of the proposed model in adjusted R^2, AIC, BIC, and nested-model tests (ANOVA).

Keywords: Heaps' Law. Quantitative Linguistics. Zipf's Law of Abbreviation.

RESUMO: A formulação clássica da Lei de Heaps estabelece uma relação bivariada entre a extensão de um texto e o crescimento de seu vocabulário, desconsiderando fatores morfológicos, fonológicos e estilísticos inerentes à linguagem natural. Este trabalho propõe e valida uma extensão compacta e generalizada da Lei de Heaps, incorporando a proporção de palavras monossílabas (p_1), fundamentada teoricamente na Lei da Abreviatura de Zipf, e uma variável indicadora de alta repetição lexical (Drep). A validação por simulação adota um processo gerador de dados (DGP) mecanístico baseado em uma urna de Zipf-Mandelbrot, no qual o vocabulário e a proporção de monossílabas emergem de um processo de amostragem de tokens, sem que a forma funcional de nenhum dos dois modelos comparados seja imposta ao processo gerador. As tabelas de resultados reportam as médias dos parâmetros estimados ao longo de 200 réplicas de Monte Carlo, garantindo que as estimativas não dependam de uma única realização amostral. A validação empírica em um grande conjunto de dados da língua inglesa, o Movie Reviews Corpus, corrobora os ganhos do modelo proposto em R^2 ajustado, AIC, BIC e testes de modelos aninhados (ANOVA).

Palavras-chave: Lei de Heaps. Linguística Quantitativa. Lei da Abreviatura de Zipf.

[1]Professora Doutora, Instituto de Matemática e Estatística - Universidade Federal de Uberlândia (IME-UFU).
[2]Doutor em Engenharia Elétrica, Faculdade de Engenharia Elétrica - Universidade Federal de Uberlândia.

RESUMEN: La formulación clásica de la Ley de Heaps establece una relación bivariada entre la extensión de un texto y el crecimiento de su vocabulario, sin considerar factores morfológicos, fonológicos y estilísticos inherentes al lenguaje natural. Este trabajo propone y valida una extensión compacta y generalizada de la Ley de Heaps, incorporando la proporción de palabras monosílabas (p_1), fundamentada teóricamente en la Ley de la Abreviación de Zipf, y una variable indicadora de alta repetición léxica (Drep). La validación por simulación adopta un proceso generador de datos (DGP) mecanicista basado en una urna de Zipf-Mandelbrot, en el cual el vocabulario y la proporción de monosílabas emergen de un proceso de muestreo de tokens, sin que la forma funcional de ninguno de los dos modelos comparados sea impuesta al proceso generador. Las tablas de resultados reportan los promedios de los parámetros estimados a lo largo de 200 réplicas de Monte Carlo, garantizando que las estimaciones no dependan de una única realización muestral. La validación empírica en un gran conjunto de datos del idioma inglés, el Movie Reviews Corpus, corrobora las ganancias del modelo propuesto en R^2 ajustado, AIC, BIC y pruebas de modelos anidados (ANOVA).

Palabras clave: Ley de Heaps. Lingüística Cuantitativa. Ley de la Abreviación de Zipf.

INTRODUCTION

The quantification and mathematical modeling of natural language constitute fundamental pillars of quantitative linguistics and textual data science. Among the most prominent statistical regularities in this field is Heaps' Law, which postulates a stable power relationship between the total length of a text, measured by the number of tokens (N), and the diversity of its vocabulary, quantified by the number of distinct types (V). Traditionally formulated as a bivariate logarithmic relationship, Heaps' Law has been widely applied to describe lexical growth across languages and textual domains.

Although formulated decades ago, classical Heaps' Law (HEAPS HS, 1978) remains, alongside Zipf's Law (ZIPF GK, 1949), the most widely used baseline empirical model in quantitative linguistics and Natural Language Processing (NLP). Its longevity is undoubtedly due to its algebraic elegance and the remarkable macroscopic approximation it offers for estimating lexical richness in large volumes of data. However, by relying exclusively on the linear length of the text, the traditional formulation disregards microstructural, phonological, and stylistic factors that inherently modulate the rate of vocabulary innovation and renewal.

Characteristics such as the density of short words (often associated with functional grammatical classes) and the degree of textual redundancy exert direct pressure on lexical expansion, elements that remain hidden in the residual noise of the bivariate model. It is precisely in this gap that the relevance of the present work lies: by proposing a compact extension that preserves the mathematical simplicity of the field's most established model

while explicitly incorporating the impact of monosyllabic density and textual redundancy, this research offers a robust correction to classical omission biases.

Several extensions and reformulations of Heaps' Law have been proposed in the quantitative linguistics literature. The classical formulation dates back to the seminal work of HERDAN G (1964) and HEAPS HS (1978), who established the power relationship between text size and observed vocabulary. More recently, DĘBOWSKI Ł (2023) proposed corrections to Heaps' Law derived from models for the hapax legomena rate (words that occur only once) as a function of text size, in a spirit similar to the present work, although with a different foundation (an urn model applied directly to hapax frequency). The asymptotic relationship between the Zipf exponent and the Heaps exponent, which theoretically underlies the mechanistic generating process used in this study, was formally derived by BAEZA-YATES R and NAVARRO G (2000), VAN LEIJENHORST DC and VAN DER WEIDE TP (2005), and LÜ L, et al. (2010). This work differs from these contributions by explicitly incorporating structural and stylistic covariates (monosyllabic density and lexical repetition) as a compact parametric extension of the classical model, and by validating this extension under a DGP that does not impose the functional form of either compared model.

METHODOLOGY

This study combines three complementary methodological components: (i) the theoretical formalization of the proposed extension of Heaps' Law, (ii) a mechanistic Monte Carlo simulation used to validate the model without circularity, and (iii) an empirical reanalysis on a real-world corpus.

Theoretical formalization of the proposed model

The traditional formulation of Heaps' Law establishes a bivariate empirical relationship between text size, measured by the total number of tokens (N), and vocabulary richness, quantified by the number of distinct types (V). In its form linearized by natural logarithms, the classical model is expressed as:

$$ln(V) = \beta_0 + \beta_1 ln(N) + \varepsilon \quad (1)$$

where $\beta_0 = ln(K)$ represents the logarithmic scale factor, β_1 the elasticity of lexical growth with respect to text length, and ε the stochastic error term. To address the analytical gap left by the bivariate formulation, the following generalized, compact extension is proposed:

$$ln(V) = \beta_0 + \beta_1 ln(N) + \beta_2 p_1 + \beta_3 Drep + \varepsilon \quad (2)$$

where p_1 is the proportion of monosyllabic words present in the vocabulary, capturing phonological density and the predominance of functional grammatical classes in the text's syntax, and Drep is a binary (dummy) control variable for lexical repetition, taking the value 1 for texts with high term redundancy and 0 for texts with a diversified distribution. β_2 and β_3 quantify, respectively, the structural impact of the monosyllabic proportion and the retraction imposed by textual redundancy on the vocabulary expansion rate.

Monte Carlo simulation design

A data-generating process (DGP) that writes the candidate model's equation directly as the simulated truth guarantees, by construction, that this model beats any more restricted alternative: this amounts to a consistency check of the estimator, not proof of superiority. For the validation to be informative, this work avoids any reduced form of $V(N)$: instead of writing the vocabulary equation directly (as in equations 1 and 2), vocabulary and the proportion of monosyllables emerge from a stochastic token sampling process drawn from a lexical frequency distribution, a Zipf-Mandelbrot urn, and neither of the two compared regression models has access to the functional form of the generating process.

A potential lexicon of $W = 8{,}000$ words is defined, indexed by a frequency rank $r = 1, \ldots, W$. The occurrence probability of each word follows the Zipf-Mandelbrot Law:

$$p(r) = \frac{(r+q)^{-s}}{\sum_r^W (r+q)^{-s}} \quad (3)$$

with a fixed Mandelbrot shift $q = 2.7$. The flattening exponent s is the only exogenous control variable introduced into the process, and it takes two values depending on document style: $s = 1.30$ for repetitive documents ($Drep = 1$) and $s = 0.95$ for diversified documents ($Drep = 0$), resting on the well-established asymptotic relationship between the Zipf exponent and the Heaps exponent ($\beta \approx 1/s \; for \; s > 1$). Each word in the lexicon is assigned, once for the entire simulation, a binary monosyllable indicator, drawn from a Bernoulli distribution whose probability decays with frequency rank, $P(monosyllable|r) =$

$logistic(2.0 - 0.70 \cdot ln\, r)$, operationally implementing Zipf's Law of Abbreviation as a structural property of the simulated language's lexicon.

For each document, a text size N is drawn (log-uniform on the interval [50,6,000] tokens), and N tokens are sampled with replacement from the distribution $p(r)$ corresponding to the document's style. The observed vocabular V is the number of distinct ranks sampled, and p_1 is the proportion of these distinct types classified as monosyllabic. Each Monte Carlo replication generates a new sample of 1,500 documents, with $Drep \sim Bernoulli(0.3)$.The simulation was replicated 200 times, with an 80/20 train/test split in each replication for computing the out-of-sample error (RMSE).

Empirical data and operationalization of variables

The empirical validation uses the Movie Reviews Corpus ($n = 2{,}000$ documents, 1,000 positive reviews and 1,000 negative reviews), originally compiled by PANG B and LEE L (2004) and widely used as a benchmark in sentiment classification tasks. For each document, the number of tokens (N) and the vocabulary (V) were obtained after removing punctuation, keeping only alphabetic tokens. The proportion of monosyllables (p_1) was calculated over the distinct types of each document, with syllable counts obtained from the CMU Pronouncing Dictionary when the word is present in the dictionary, and by a vowel-nucleus counting heuristic as an alternative for words outside the dictionary (proper nouns, slang, typos).

The variable Drep was operationalized using Yule's K (YULE GU, 1944), a classical measure of vocabulary concentration widely used in stylometry:

$$K = 10^4 \times [\Sigma_i i^2 V_i - N]/N^2 \quad (4)$$

where V_i is the number of types occurring exactly i times in the document. Yule's K is high when a few words account for a large fraction of occurrences (more repetitive text) and low when occurrences are more uniformly distributed across many distinct types (more diverse text). Instead of an arbitrary threshold such as the tercile, the Drep classification was obtained through unsupervised k-means clustering ($k = 2$) applied directly to the corpus's one-dimensional vector of Yule's K values (MACQUEEN J, 1967; LLOYD S, 1982), following the validation approach proposed by LINHARES RR and VARGAS RN (2026): the algorithm partitions documents into two groups, assigning each to the nearest centroid (average K),

without requiring an a priori cutoff point; the cluster with the higher centroid was labeled as high lexical repetition ($Drep = 1$).

RESULTS

Across 200 Monte Carlo replications under the mechanistic Zipf-Mandelbrot DGP, all coefficients common to both models (intercept and $ln(N)$) were significant at 5% in 100% of the replications, as were the coefficients of p_1 and $Drep$ in the proposed model. Table 1 reports the average estimated coefficients (with average standard errors and t statistics in parentheses and brackets), and Table 2 reports the average fit and out-of-sample predictive metrics for both models.

The empirical reanalysis on the Movie Reviews Corpus shows that the classical bivariate Heaps' Law model already achieves an adjusted R^2 = 0.9525, with $\beta_1 = 0.7619$ ($p < 2 \times 10^{-16}$) and a residual standard error of 0.0763. Incorporating p_1 and $Drep$ into the proposed model raised the adjusted R^2 to 0.9683, with a residual standard error of 0.0623. Of the 2,000 documents in the corpus, the k-means procedure classified 724 (36.2%) as $Drep = 1$, with K centroids of 83.16 (low-repetition group) and 112.07 (high-repetition group). Table 3 reports the estimated coefficients of the proposed model. The superiority of the proposed model was confirmed by the F test for nested models ($F = 500.8$; $p < 2.2 \times 10^{-16}$), and by the reductions in AIC (from −4,613.47 to −5,422.82) and BIC (from −4,602.26 to −5,400.41), summarized in Table4.

Variable / Metric	Classical Model (average estimate)	Proposed Model (average estimate)
Intercept (β_0)	0.7388 (0.0344) [t = 21.45]	0.6068 (0.0365) [t = 16.66]
$ln(N)(\beta_1)$	0.7932 (0.0053) [t = 148.79]	0.8234 (0.0040) [t = 205.69]
p_1 (Monosyllables)	N/A	0.7290 (0.0867) [t = 8.42]
$Drep$ (Repetition)	N/A	−0.5790 (0.0082) [t = −70.43]

Table 1: Average estimated coefficients on data generated by the Zipf-Mandelbrot urn (average of 200 Monte Carlo replications, n = 1,500 documents per replication). All reported coefficients were significant at 5% in 100% of the replications. Source: prepared by the authors (2026).

Variable / Metric	Classical Model (average estimate)	Proposed Model (average estimate)
Adjusted R^2 (average)	0.9485	0.9947
AIC (average)	133.0	−2582.6
BIC (average)	148.2	−2557.2
Out-of-sample RMSE (average)	0.2560	0.0821
% of replications in which BIC prefers the model	N/A	100%
% of replications in which OOS RMSE prefers the model	N/A	100%

Table 2: Average fit and predictive-power metrics on data generated by the Zipf-Mandelbrot urn (average of 200 Monte Carlo replications, n = 1,500 documents per replication). Source: prepared by the authors (2026).

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept (β_0)	1.6004	0.0332	48.17	$< 2\times10^{-16}$
$ln(N)$	0.7148	0.0036	200.00	$< 2\times10^{-16}$
p_1 (Monosyllables)	−0.7960	0.0303	−26.27	$< 2\times10^{-16}$
$Drep$ (Yule's K + K-means)	−0.0533	0.0029	−18.37	$< 2\times10^{-16}$

Table 3: Estimated coefficients of the proposed model (Movie Reviews Corpus, reanalysis, n = 2,000, Drep via Yule's K + K-means). Source: prepared by the authors (2026), based on PANG and LEE (2004).

Metric	Classical Model	Proposed Model
Adjusted R^2	0.9525	0.9683
Residual Std. Error	0.0763	0.0623
Degrees of Freedom (df)	1998	1996
Residual Sum of Squares (RSS)	11.6382	7.7494
AIC	−4613.47	−5422.82
BIC	−4602.26	−5400.41

Table 4: Fit comparison between the models (Movie Reviews Corpus, reanalysis, n = 2,000, Drep via Yule's K + K-means). Source: prepared by the authors (2026).

Figure 1 shows the scatter of the 2,000 documents in the Movie Reviews Corpus with respect to text size (N) and vocabulary size (V), overlaying the curves fitted by the classical model and by the proposed model, the latter evaluated at the sample averages of p_1 and $Drep$, so as to produce an average-profile curve comparable to that of the classical model.

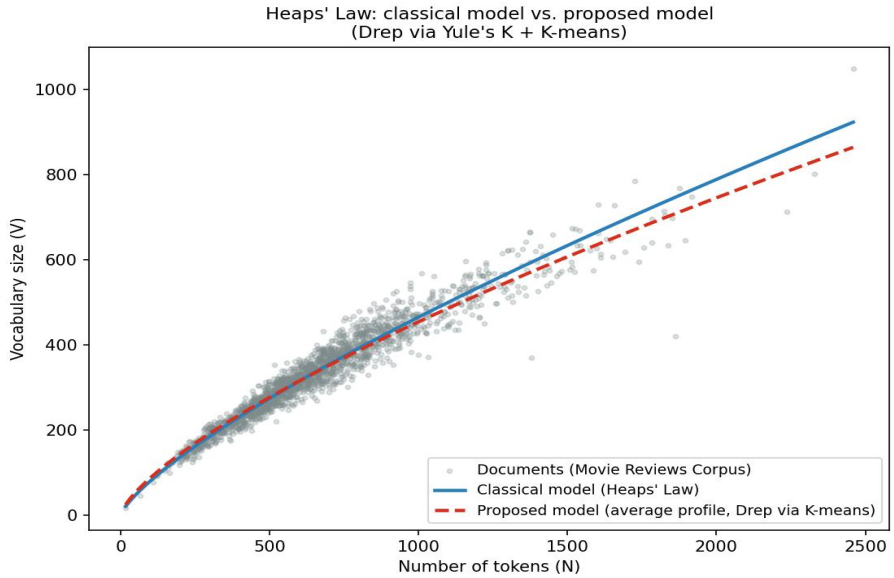

Figure 1: *Comparison between the classical Heaps' Law model and the proposed model (average profile, Drep via Yule's K + K-means) applied to the Movie Reviews Corpus (reanalysis, n = 2,000). Source: prepared by the authors (2026).*

Overall, the proposed extension performed well across both validation exercises. In the mechanistic Monte Carlo simulations, it consistently outperformed the classical model in both within-sample fit and out-of-sample prediction, with the coefficients of p_1 and Drep significant in 100% of the replications. In the empirical reanalysis of the Movie Reviews Corpus, it delivered analogous gains in adjusted R^2, AIC, BIC, and the F test for nested models, with both covariates again significant and stable. This convergence between simulated and real-world evidence indicates that the improvement is not an artifact of either exercise in isolation, but a robust property of the proposed compact extension of Heaps' Law.

CONCLUSION

This research addressed the structural limitations of the classical formulation of Heaps' Law, proposing a compact extension that incorporates the proportion of monosyllabic words (p_1) and a variable indicating high lexical repetition ($Drep$). The simulation-based validation adopts a mechanistic generating process based on a Zipf-Mandelbrot urn, in which vocabulary emerges from a token sampling process independent of both candidate models, avoiding the circularity present in Monte Carlo designs that directly impose the functional form of the proposed model. Across 200 Monte Carlo replications, the proposed model reduced the out-of-sample prediction error by approximately 68% relative to the classical model, with coefficients of p_1 and $Drep$ significant at 5% in 100% of the replications. This evidence is complemented

by the empirical reanalysis on the Movie Reviews Corpus, which corroborates the gains in adjusted R^2, AIC, BIC, and the F test for nested models, and by the graphical comparison of the curves fitted by the two models.

As future developments, we suggest: (a) applying the compact extension to other morphologically distinct languages; (b) investigating interactions between p_1, Drep, and register/genre indicators across corpora from varied domains; (c) formally testing the bimodality of the Yule's K distribution underlying the k-means classification of $Drep$; and (d) exploring even more realistic mechanistic generating processes (for example, Pitman-Yor processes with a frequency base fitted to real corpora) as a validation standard for future extensions of Heaps' Law.

REFERENCES

BAEZA-YATES R, NAVARRO G. Block addressing indices for approximate text retrieval. Journal of the American Society for Information Science, 2000.

DĘBOWSKI Ł. Corrections of Zipf's and Heaps' Laws Derived from Hapax Rate Models. arXiv:2307.12896, 2023.

HEAPS HS. Information Retrieval: Computational and Theoretical Aspects. Academic Press, 1978.

HERDAN G. Quantitative Linguistics. Butterworths, 1964.

LINHARES RR, VARGAS RN. Separating documents by lexical repetition via Yule's K and K-means: validation through Monte Carlo simulation. Revista Ibero-Americana de Humanidades, Ciências e Educação, 2026; 12(9): 1-9.

LLOYD S. Least squares quantization in PCM. IEEE Transactions on Information Theory, 1982; 28(2): 129-137.

LÜ, L.; ZHANG, Z.-K.; ZHOU, T. Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems. PLoS ONE, 2010; 5(12): e14139.

MACQUEEN J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967; 1: 281-297.

PANG B, LEE L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004; p. 271-278.

VAN LEIJENHORST DC, VAN DER WEIDE TP. A formal derivation of Heaps' Law. Information Sciences, 2005; 170(2-4): 263-272.

YULE GU. The Statistical Study of Literary Vocabulary. Cambridge University Press, 1944.

ZIPF, G. K. Human Behavior and the Principle of Least Effort. Cambridge, MA: Addison-Wesley, 1949.