# A Model for Predicting $n$-gram Frequency Distribution in Large *Corpora*

Joaquim F. Silva[1[0000−0002−5223−1180]] and Jose C. Cunha[1[0000−0001−6729−8348]]

NOVA Laboratory for Computer Science and Informatics {jfs,jcc}@fct.unl.pt**

**Abstract.** The statistical extraction of multiwords ($n$-grams) from natural language *corpora* is challenged by computationally heavy searching and indexing, which can be improved by low error prediction of the $n$-gram frequency distributions. For different $n$-gram sizes ($n \geq 1$), we model the sizes of groups of equal-frequency $n$-grams, for the low frequencies, $k = 1, 2, \ldots$, by predicting the influence of the *corpus* size upon the Zipf's law exponent and the $n$-gram group size. The average relative errors of the model predictions, from 1-grams up to 6-grams, are near 4 %, for English and French *corpora* from 62 Million to 8.6 Billion words.

**Keywords:** $n$-gram frequency distribution · large corpora.

## 1 Introduction

Relevant Expressions (RE) are semantically meaningful $n$-grams ($n \geq 1$), as "oceanography", "oil crisis", useful in document classification [15] and $n$-gram applications. However, most word sequences are not relevant in a *corpus*. Statistical RE extraction from texts, e.g [18, 7], measures the cohesion among the $n$-grams within each distinct multiword; its performance benefits from predicting the $n$-gram frequency distributions. Low frequency $n$-grams are significant proportions of the number of distinct $n$-grams in a text, as well as of the RE. Assuming, for language $L$ and $n$-gram size $n$, a finite vocabulary $V(L, n)$ in each temporal epoch [17, 9, 16], we model the influence of the corpus size upon the sizes $W(k)$ of equal-frequency ($k$) $n$-gram groups, for $n \geq 1$, especially for low frequencies. We present results (and compare to a Poisson-based model), for English and French Wikipedia *corpora* (up to 8.6 Gw), for $1 \leq n \leq 6$. We discuss background, the model, results and conclusions.

## 2 Background

Zipf's law [20] is a good approximation to word frequency distribution, deviating from real data in high and low frequencies. More accurate approximations pose open issues [12, 19, 2, 1, 11, 8, 6, 13, 14]. Low frequency words are often ignored,

** Acknowledgements to FCT MCTES, NOVA LINCS UIDB/04516/2020 and Carlos Gonçalves.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_55

as well as multiwords. Most studies use truncated *corpora* data [5, 8], with some exceptions [17]. In models as [2, 3] the probability of a word occurring $k$ times is given by a power law $k^{-\gamma}$ corrected by the *corpus* size influence, but they do not consider other $n$-gram sizes, unlike e.g. [16].

## 3 The Model

Successive model refinements are shown: $W_z(k)$, from Zipf's Law; $W_{\alpha_d}(k,C)$ for *corpus* size dependence; and $W^*(k, C)$ for scaling adjustments.

### 3.1 $W_z(k)$: The Size of the Frequency Levels from Zipf's Law

By Zipf's Law [20], the number of occurrences of the $r^{\text{th}}$ most frequent word in a *corpus* with a number of distinct words given by $D$, is

$$f(r) = f(1) \cdot r^{-\alpha} \ , \tag{1}$$

$\alpha$ is a constant $\sim 1$; $r$ is the word rank ($1 \leq r \leq D$). (1) also applies to $n$-grams of sizes $n > 1$, with $\alpha$ dependent on $n$ (for simplicity $\alpha$ replaces $\alpha(n)$). The relative frequency of the most frequent $n$-gram ($r = 1$) for each $n$ shows small fluctuations around a value, taken as an approximation to its occurrence probability, $p_1$. The absolute frequency $f(1) \approx p_1 \cdot C$. So, $\ln(f(r))$ would decrease linearly with slope $\alpha$ as $\ln(r)$ increases. Real distributions deviate from straight lines and show, for their higher ranks, groups of equal-frequency $n$-grams. $W(k)$ is defined based on Zipf's law [4, 16]. For a level with frequency $k$, with its lowest ($r_{l_k}$) and highest ($r_{h_k}$) $n$-gram ranks: $f(r_{l_k}) = f(r_{h_k}) = k$; $W_z(k) = r_{h_k} - r_{l_k} + 1$. The model assumes a minimum observed frequency of 1: $f(r_{l_1}) = f(r_{h_1}) = 1$; $r_{h_1} = D$; and only applies to the higher ranks / lower frequencies where adjacent levels ($r_{l_k} = r_{h_{k+1}} + 1$) have consecutive integer frequencies: $f(r_{h_{k+1}}) = f(r_{h_k}) + 1$. Then, (2) is obtained, with constant $\alpha_z$.

$$W_z(k) = \left(\frac{1}{D^{\alpha_z}} + \frac{k-1}{f(1)}\right)^{-\frac{1}{\alpha_z}} - \left(\frac{1}{D^{\alpha_z}} + \frac{k}{f(1)}\right)^{-\frac{1}{\alpha_z}} \ . \tag{2}$$

$$D(C; L, n) = \frac{V(L, n)}{1 + (K_2 \cdot C)^{-K_1}} \ . \tag{3}$$

For predicting $D$ in a *corpus* of size $C$, we use (3), following [16] with good agreement with real *corpora*. For language $L$ and $n$-gram size $n$, $V(L, n)$ is the finite vocabulary size; $K_1$, $K_2$ are positive constants. If $V$ is assumed infinite, (3) equals Heap's law.

### 3.2 An Analytical Model for the Dependence of $\alpha$ on *Corpus* Size

Empirically, $\alpha_z$ is shown to depend on *corpus* size. So, we consider $\alpha$ in (1) as a function $\alpha(C, r)$ of the *corpus* size and the $n$-gram rank $r$:

$$\alpha(C, r) = \frac{\ln(f_c(1)) - \ln(f_c(r))}{\ln(r)} \ , \tag{4}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_55

where $1 \leq r \leq D$, and $f_c(1)$ and $f_c(r)$ are the frequencies, respectively, of the most frequent $n$-gram and the r$^{\text{th}}$ ranked $n$-gram, in a *corpus* of size $C$. In (2) $\alpha$ is obtained, for each *corpus* size, by fitting $W_z(1)$ to the empirical level size $W_{obs}(1)$ (for $k=1$). For that level, $r_{h_1} = D(C, L, n)$ (denoted $D$ or $D_c$), and $f_c(D_c) = 1$, so $\ln(f_c(r)) = 0$ in (4) for $r = D_c$. Let $\alpha(C, D_c)$ (denoted $\alpha_d(C)$), be the $\alpha$ value at rank $D$. Let $C_1$ be the size of a reference *corpus*:

$$\alpha_d(C) - \alpha_d(C_1) = \frac{\ln(f_c(1))}{\ln(D_c)} - Ref_{c_1} \ . \tag{5}$$

The 2$^{\text{nd}}$ term in the right-hand side of (5) (denoted $Ref_{c_1}$) becomes fixed. It only depends on $f_c(1) = C_1 \cdot p_1$ ($p_1$ is the occurrence probability of the most frequent $n$-gram) and $D_{c_1}$ from (3). Using Table 1 (Section 4.2) and tuning $\alpha_d(C_1)$ by fitting, for $C_1$, the $W_z(1)$ from (2) to the observed $W_{obs}(1)$, we find $\alpha_d(C_1)$ and $D_{c_1}$. Given $\alpha_d(C_1)$ and $Ref_{c_1}$, then (5) predicts $\alpha_d(C)$ for a size $C$ *corpus*, and $W_z(k)$ (2) leads to $W_{\alpha_d}(k,C)$ (6), where $\alpha_d(C)$ replaces $\alpha_z$:

$$W_{\alpha_d}(k,C) = \left(\frac{1}{D_c^{\alpha_d(C)}} + \frac{k-1}{f_c(1)}\right)^{-\frac{1}{\alpha_d(C)}} - \left(\frac{1}{D_c^{\alpha_d(C)}} + \frac{k}{f_c(1)}\right)^{-\frac{1}{\alpha_d(C)}} \ . \tag{6}$$

### 3.3 $W^*(k, C)$: The Dependence of Level Size on *Corpus* Size

The frequency level size depends on frequency $k$ and *corpus* size $C$. Firstly, for a *corpus* size $C$, $\alpha_z$ in (2) is tuned to best fitting $W_z(1)$ to $W_{obs}(1)$. Except for the $W_{obs}(k)$ fluctuations (Fig. 1a), the deviation, closely proportional to $\ln(k)$, between $W_{obs}(k)$ and $W_z(k)$, suggests the improvements due to (7) (Fig. 1a).

$$W_{adjusted}(k) = W_z(k) \cdot k^{\beta} \ . \tag{7}$$

$\beta$ is a constant for each $n$, obtained from the best fit of $W(k)$ to $W_{obs}(k)$, for a given *corpus*. Secondly, for different *corpus* sizes, Fig. 1b shows $W_{obs}(k)$ curves as a function of $k$, seeming parallel, but a detailed analysis shows otherwise. If, for each $\ln(W_{obs}(k, C^*))$ for the three smaller *corpora* $C^*$, an offset equal to $\ln(W_{obs}(1, C)) - \ln(W_{obs}(1, C^*))$ is added ($C = 8.6$ Gw being the largest *corpus*), the resulting curves (omitted due to lack of space) do not coincide, as they should if they were parallel in Fig. 1b. The gap between the curves is proportional to $\ln(k)$. And, for each $\ln(k)$ value, the distance in $\ln(W_{obs}(k))$ for *corpora* of sizes $C$ and $C_1$ is proportional to $\log_2(C/C_1)$. The distance between the $\ln(W(k))$ curves of any two *corpora* $C$ and $C_1$ is approximated by (8), with $\delta$ constant for each $n$. Joining (6), (7), (8) leads to the final model, $W^*(k, C)$, (9):

$$\Delta = \delta \cdot \ln(\frac{C}{C_1}) \cdot \ln(k) \tag{8}$$

$$W^*(k, C) = W_{\alpha_d}(k, C) \cdot k^{\beta + \delta \cdot \ln(\frac{C}{C_1})} \ . \tag{9}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_55

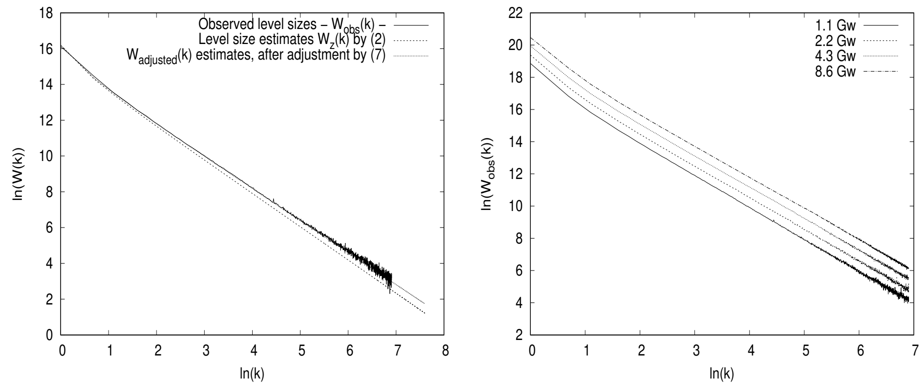


Fig. 1: a) 1-gram equal-frequency level size $W(k)$ *vs* $k$ (log-log scale) – observed and estimates by (2) and (7) from a 1.1 Gw English *corpus*; b) Observed 3-gram level size values, $W_{obs}(k)$ *vs* $k$ (log-log scale), for different English *corpora* sizes.

# 4 Results and Discussion

## 4.1 The Poisson-Zipf Model

In the $W_P(k, C)$ model of [17] given by (10), an $n$-gram ranked $r$ occurs, in a size $C$ *corpus*, a number of times following Poisson distribution [10] with $\lambda_r = f(r)$ by Zipf's Law. $W(0) = W_P(0, C)$ is the estimated number of unseen $n$-grams in the *corpus*. $D = V - W(0)$, for $n$-gram vocabulary size $V$.

$$W_P(k, C) = \sum_{r=1}^{r=V} \frac{(p_1 \cdot C \cdot r^{-\alpha})^k \cdot e^{-p_1 \cdot C \cdot r^{-\alpha}}}{k!} \approx \int_1^V \frac{(p_1 \cdot C \cdot r^{-\alpha})^k \cdot e^{-p_1 \cdot C \cdot r^{-\alpha}}}{k!} dr$$
$$\approx \frac{(p_1 \cdot C)^{1/\alpha}}{\alpha \cdot k!} \cdot \left[ \Gamma(k - \frac{1}{\alpha}, \frac{p_1 \cdot C}{V^\alpha}) - \Gamma(k - \frac{1}{\alpha}, p_1 \cdot C) \right] \quad (10)$$

## 4.2 Comparison of Results

Complete *corpora* were built from documents randomly extracted from English and French Wikipedia. For evaluating size dependence, they were doubled successively (Table 2). A space was added between the words and each of the following characters: {!, ?, :, ;, ,, (, ), [, ], <, >, "}. All inflected word forms were kept.

***The Model Calculations.*** (I) To calculate $D(C; L, n)$ in (3), parameters $K_1$, $K_2$ and $V(L, n)$ were found for each language $L$ and $n$-gram size $n$ (Table 1, also showing the $\beta$ and $\delta$ values used in (9)). The $V(L, n)$ value is an estimate of the vocabulary size, such that further increasing it, does not significantly reduce the relative error $((E - O)/O) \cdot 100\%$, between an estimated value ($E$) and the corresponding observed value ($O$). Pairs ($K_1$, $K_2$) were found leading

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_55

to the lowest possible relative error, for a selected pair of *corpora* with sizes close to the lowest and highest *corpora* sizes in the considered range for each language. (II) To evaluate the relative errors, (9) was applied with $k$ such that the observed level sizes of consecutive frequency levels $k$ and $k+1$ are monotonic decreasing, $W_{obs}(k,C) > W_{obs}(k+1,C)$. This avoids the non-monotony regions of the observed $\ln(W(k))$ curve (Fig. 1a). We considered a basic set of $k$ values, $K = \{1,2,3,4,5,6,7,8,16,32,64,128\}$, constrained (to ensure $\ln(W(k))$ monotony) depending on the *corpus* size $C$: for $C < 250\,\mathrm{Mw}$, we used $k \leq 16$; for $C < 1\,\mathrm{Gw}$, $k \leq 32$; the full $K$ set was used only for $C > 4\mathrm{Gw}$. We selected *corpora* of sizes ($C_1$) 1.1 Gw (English) and 808 Mw (French). (III) The $\alpha_d(C_1)$ values for $n$-gram sizes from 1 to 6 are: (English) 1.1595, 1.02029, 0.88825, 0.82532, 0.8117, 0.8027; (French) 1.158825, 1.0203, 0.86605, 0.84275, 0.80818, 0.7569. The empirical values of $p_1$ for $n$-gram sizes from 1 to 6: (English) 0.06704, 0.03250, 0.0062557, 0.0023395, 0.0017908, 0.0014424; (French) 0.07818, 0.037976, 0.004685, 0.0036897, 0.001971, 0.00072944. (IV) To run $W_P(K,C)$, $\alpha$ values leading to the lowest relative errors, are, for $n$-gram sizes from 1 to 6: (English) 1.17, 1.02, 0.891, 0.827, 0.814, 0.812; (French) 1.156, 1.01, 0.884, 0.842, 0.806, 0.759.

Table 1: Parameter values $K_1$, $K_2$ and vocabulary sizes ($V(L,n)$) to be used in $D(C;L,n)$, (3), and $\beta$ and $\delta$ to $W^*(k,C)$, (9).

| | **English** | | | | | |
|---|---|---|---|---|---|---|
| | **1-grams** | **2-grams** | **3-grams** | **4-grams** | **5-grams** | **6-grams** |
| $\boldsymbol{K_1}$ | 0.838 | 0.861 | 0.885 | 0.924 | 0.938 | 0.955 |
| $\boldsymbol{K_2}$ | 3.61 e−11 | 5.1 e−11 | 2.66 e−11 | 1.78 e−11 | 4.29 e−12 | 6.5 e−13 |
| $\boldsymbol{V}$ | 2.45 e+8 | 9.9 e+8 | 4.74 e+9 | 1.31 e+10 | 6.83 e+10 | 5.29 e+11 |
| $\boldsymbol{\beta}$ | 0.044 | 0.113 | 0.129 | 0.135 | 0.122 | 0.082 |
| $\boldsymbol{\delta}$ | 0.0039 | 0.0118 | 0.0310 | 0.0353 | 0.0339 | 0.0331 |
| | **French** | | | | | |
| | **1-grams** | **2-grams** | **3-grams** | **4-grams** | **5-grams** | **6-grams** |
| $\boldsymbol{K_1}$ | 0.809 | 0.794 | 0.838 | 0.867 | 0.903 | 0.907 |
| $\boldsymbol{K_2}$ | 4.501 e−11 | 3.801 e−11 | 3.901 e−11 | 2.501 e−11 | 2.201 e−11 | 2.01 e−12 |
| $\boldsymbol{V}$ | 2.35 e+8 | 1.095 e+9 | 3.1 e+9 | 8.18 e+9 | 1.41 e+10 | 1.45 e+11 |
| $\boldsymbol{\beta}$ | 0.0812 | 0.120 | 0.175 | 0.140 | 0.160 | 0.234 |
| $\boldsymbol{\delta}$ | 0.0061 | 0.0190 | 0.0354 | 0.0491 | 0.0469 | 0.0384 |

Table 2 presents the relative errors for the predictions of the frequency level sizes. For each $n$-gram size, the left column refers to $W^*(k,C)$ and the right one to $W_P(k,C)$. For each pair (*corpus* size, $n$-gram size), it shows the *average relative error* for the $K$ set used: $AvgErr(K) = \frac{1}{\|K\|}\sum_{k\in K} Err(k)$, where $Err(k) = |\frac{W(k,C)-W_{obs}(k,C)}{W_{obs}(k,C)}|$. The *average relative errors* for $W^*(k,C)$ are much lower than for $W_P(k,C)$, which assumes an ideal Zipf's Law. The line **Avg** shows the average value of each column over the full range of *corpora* sizes, with errors of the same magnitude across the range of $n$-gram sizes for $W^*(k,C)$, but having significant variations in the **Avg** values for $W_P(k,C)$. The *global relative error*

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_55

is the average of the **Avg** values over the range of $n$-gram sizes, being around 4 % for $W^*(k, C)$. Thus, $W^*(k, C)$ curves (omitted due to lack of space) closely follow the $W_{obs}(k, C)$ curves forms of Fig. 1.

Table 2: Average relative error (%) for the predictions of the $n$-gram frequency level sizes obtained by $W^*(k, C)$, (9), (left col.), and $W_P(k, C)$, (10), (right col.). Each cell in the table gives an average relative error over a subset of $k$ values within the set $K$ considered for that cell, as described in the text.

| | | | | | | English | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Corpus*** | **1-grams** | | **2-grams** | | **3-grams** | | **4-grams** | | **5-grams** | | **6-grams** | |
| 63 Mw | 5.1 | 29.9 | 7.5 | 42.0 | 3.1 | 50.1 | 4.4 | 53.7 | 6.3 | 54.5 | 5.8 | 64.7 |
| 128 Mw | 2.2 | 23.7 | 2.8 | 32.4 | 4.4 | 42.5 | 6.6 | 45.9 | 5.7 | 49.5 | 7.1 | 60.3 |
| 255 Mw | 3.1 | 22.1 | 2.0 | 28.4 | 2.6 | 36.3 | 5.9 | 37.6 | 3.7 | 40.0 | 5.4 | 49.8 |
| 509 Mw | 4.9 | 15.8 | 4.7 | 22.2 | 3.7 | 29.4 | 4.6 | 30.1 | 4.7 | 31.9 | 6.6 | 40.3 |
| 1.1 Gw | 3.1 | 13.3 | 2.6 | 19.8 | 3.4 | 26.4 | 3.9 | 26.8 | 5.4 | 27.9 | 5.2 | 32.9 |
| 2.2 Gw | 5.1 | 9.5 | 6.2 | 23.2 | 3.9 | 26.7 | 3.3 | 28.5 | 4.9 | 31.5 | 3.7 | 28.7 |
| 4.3 Gw | 2.8 | 10.7 | 2.7 | 28.5 | 2.3 | 34.8 | 3.1 | 37.4 | 4.2 | 39.3 | 4.8 | 31.4 |
| 8.6 Gw | 6.1 | 13.4 | 6.7 | 37.6 | 4.4 | 47.2 | 6.0 | 51.7 | 5.9 | 52.4 | 6.5 | 40.4 |
| **Avg** | **4.1** | 17.3 | **4.4** | 29.3 | **3.5** | 36.7 | **4.7** | 39.0 | **5.1** | 40.9 | **5.6** | 43.6 |
| | | | | | | **French** | | | | | | |
| ***Corpus*** | **1-grams** | | **2-grams** | | **3-grams** | | **4-grams** | | **5-grams** | | **6-grams** | |
| 108 Mw | 2.8 | 22.6 | 2.4 | 30.9 | 2.6 | 54.9 | 4.2 | 40.8 | 4.9 | 42.3 | 5.5 | 65.4 |
| 201 Mw | 1.5 | 18.6 | 2.0 | 25.7 | 2.1 | 49.8 | 2.9 | 33.0 | 3.2 | 33.9 | 3.5 | 57.7 |
| 404 Mw | 2.7 | 15.5 | 2.9 | 23.5 | 4.0 | 44.4 | 4.6 | 28.4 | 4.9 | 29.4 | 5.0 | 51.6 |
| 808 Mw | 2.9 | 12.6 | 3.0 | 26.7 | 3.2 | 34.8 | 3.4 | 23.6 | 3.6 | 24.3 | 3.7 | 43.8 |
| 1.61 Gw | 4.6 | 16.9 | 3.5 | 37.2 | 3.0 | 29.0 | 2.9 | 28.7 | 3.3 | 29.6 | 3.4 | 41.9 |
| 3.2 Gw | 4.0 | 19.2 | 3.2 | 48.6 | 4.2 | 22.8 | 5.3 | 39.5 | 3.3 | 41.1 | 6.6 | 49.7 |
| **Avg** | **3.1** | 17.6 | **2.8** | 32.1 | **3.7** | 39.3 | **3.9** | 32.3 | **3.9** | 33.4 | **4.6** | 51.7 |

## 5 Conclusions

Estimating $n$-gram frequency distributions is useful in statistical-based $n$-gram applications. The proposed model estimates the sizes $W(k, C)$ of equal-frequency ($k$) $n$-gram groups in a *corpus* of size $C$, for the low frequency $n$-grams. It applies uniformly to different $n$-gram sizes $n \geq 1$ and languages, assuming a finite language $n$-gram vocabulary. It models the dependences of Zipf's Law exponent and $W(k, C)$ on $C$, agreeing well with $n$-gram frequency data from unigrams up to hexagrams, from real un-truncated English and French *corpora* with million to billion words. Larger *corpora* evaluation is planned.

## References

1. Ausloos, M., Cerqueti, R.: A universal rank-size law. PLoS ONE **11**(11) (2016)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_55

2. Balasubrahmanyan, V.K., Naranan, S.: Algorithmic information, complexity and zipf's law. Glottometrics **4**, 1–26 (2002)
3. Bernhardsson, S., da Rocha, L.E.C., Minnhagen, P.: The meta book and size-dependent properties of written language. New Journal of Physics **11**(12) (2009)
4. Booth, A.D.: A law of occurrences for words of low frequency. "Inform. & Control" **10**(4), 386–393 (1967)
5. Brants, T., Popat, A.C., Xu, P., Och, F.J., Dean, J.: Large language models in machine translation. In: Joint Conf. on Empirical Methods in NLP and Computational Natural Language Learning. pp. 858–867. ACL (2007)
6. Cancho, R.F., Solé, R.V.: Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited*. Journal of Quantitative Linguistics **8**(3), 165–173 (2001)
7. Dias, G.: Multiword unit hybrid extraction. In: ACL Workshop on Multiword Expressions, Vol. 18. pp. 41–48. ACL (2003)
8. Gerlach, M., Altmann, E.G.: Stochastic model for the vocabulary growth in natural languages. Phys. Rev. X **3**, 021006 (May 2013)
9. Goncalves, C., Silva, J., Cunha, J.C.: n-gram cache performance in statistical extraction of relevant terms in large corpora. In: et al., J.R. (ed.) Proceedings of the ICCS 2019. LNCS, vol. 11537, pp. 75–88. ICCS, Springer, Algarve, Portugal (2019)
10. Haight, F.A.: Handbook of the Poisson Distribution. John Wiley & Sons, New York (1967)
11. Lü, L., Zhang, Z.K., Zhou, T.: Deviation of zipf's and heaps' laws in human languages with limited dictionary sizes. Scientific Reports **3**(1082) (2013)
12. Mandelbrot, B.: On the theory of word frequencies and on related markovian models of discourse. In: Struct. of Language and its Mathematical Aspects (1953)
13. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. Internet Math. **1**(2), 226–251 (2003)
14. Piantadosi, S.T.: Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review **21**, 1112–1130 (2014)
15. Silva, J., Mexia, J., Coelho, A., Lopes, G.: Document clustering and cluster topic extraction in multilingual corpora. In: Proceedings 2001 IEEE International Conference on Data Mining. pp. 513–520 (2001)
16. Silva, J.F., Cunha, J.C.: An empirical model for n-gram frequency distribution in large corpora. In: Lauw, H.W., Wong, R.C.W., Ntoulas, A., Lim, E.P., Ng, S.K., Pan, S.J. (eds.) Advances in Knowledge Discovery and Data Mining, 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II. Springer International Publishing (2020)
17. Silva, J.F., Gonçalves, C., Cunha, J.C.: A theoretical model for n-gram distribution in big data corpora. In: 2016 IEEE Intl. Conf. on Big Data. pp. 134–141 (2016)
18. Silva, J.F., Dias, G., Guilloré, S., Pereira Lopes, J.G.: Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In: Barahona, P., Alferes, J.J. (eds.) Progress in Artificial Intelligence. pp. 113–132. Springer Berlin Heidelberg, Berlin, Heidelberg (1999)
19. Simon, H.: On a class of skew distribution functions. Biometrika **42**(3/4), 425—-440 (1955)
20. Zipf, G.K.: Human Behavior and the Principle of Least-Effort. Addison-Wesley, Cambridge, MA (1949)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_55