An Empirical Model for n-gram Frequency Distribution in Large Corpora

Joaquim F. Silva(✉) and Jose C. Cunha

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

Abstract. Statistical multiword extraction methods can benefit from the knowledge on the n-gram ($n \geq 1$) frequency distribution in natural language corpora, for indexing and time/space optimization purposes. The appearance of increasingly large corpora raises new challenges on the investigation of the large scale behavior of the n-gram frequency distributions, not typically emerging on small scale corpora. We propose an empirical model, based on the assumption of finite n-gram language vocabularies, to estimate the number of distinct n-grams in large corpora, as well as the sizes of the equal-frequency n-gram groups, which occur in the lower frequencies starting from 1. The model was validated for n-grams with $1 \leq n \leq 6$, by a wide range of real corpora in English and French, from 60 million up to 8 billion words. These are full non-truncated corpora data, that is, their associated frequency data include the entire range of observed n-gram frequencies, from 1 up to the maximum. The model predicts the monotonic growth of the numbers of distinct n-grams until reaching asymptotic plateaux when the corpus size grows to infinity. It also predicts the non-monotonicity of the sizes of the equal-frequency n-gram groups as a function of the corpus size.

Keywords: n-gram frequency distribution · Large text *corpora*.

1 Introduction

The appearance of Web-scale *corpora* raised new challenges on the extraction of relevant expressions in natural languages, e.g. for indexing and time/space optimization, whose efficiency can benefit from the knowledge of the statistical regularities in real data. However, most studies only focus on single words, analyzing their occurrence frequencies. For example, function words such as "the", "in", "of", lacking semantic content and having a small and fixed vocabulary essentially related to a language grammar, tend to occur more often than words like "oceanography" or "preferably", whose appearance can be related to the semantic content of a text.

These studies should be extended with more generic approaches for the extraction of multiword expressions based on the properties of n-grams. An n-gram is a

Acknowledgements to FCT MCTES, NOVA LINCS UID/CEC/04516/2019 and Carlos Gonçalves.

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12085, pp. 840–851, 2020.
https://doi.org/10.1007/978-3-030-47436-2_63

sequence of $n \geq 1$ consecutive words, so, beyond single words, its characteristics can be related to the text phrases and sentences, e.g. "History of Science". In a given *corpus* one can observe distinct n-gram types, each one showing a certain number of instances. This requires an accurate estimation of the n-gram frequency distributions for any given *corpus* size, particularly important in Big Data extraction applications handling many Mega (10^6) and Giga (10^9) words.

We present a model that estimates, with good accuracy, the total numbers of distinct n-grams ($1 \leq n$) in real *corpora* for a wide range of sizes, in a given language. It also estimates the sizes of the frequency levels, i.e. the numbers of equal-frequency n-grams, for the extreme low frequencies, from the singletons onwards. The lower frequency n-grams are a significant proportion of the distinct n-grams across a wide range of *corpora* sizes, and a large part of the relevant expressions in a text. The model predicts the finite sizes of the n-gram vocabularies in a given language, from 1-grams to 6-grams. This range of n-gram sizes captures the most meaningful relevant expressions. It also predicts growth of the population of distinct n-grams towards asymptotic plateaux, for large enough *corpus*. The model also predicts that, for the lowest frequencies, the numbers of distinct n-grams with equal frequencies, instead of always growing with the *corpus* size, will present a non-monotonic behavior, reaching a maximum and then decreasing as *corpus* grows to infinity. Results were validated with full non-truncated data from English and French Wikipedia *corpora* from 60 Mega to 8 Giga words. We discuss background (Sect. 2), the model (Sect. 3), the results (Sect. 4) and conclusions.

2 Background

The empirical Zipf's Law [11] is a known model for word frequency distributions with a power law approximation in good agreement with data in a large range of frequencies, but significant deviations in the high and low frequencies. Most studies recognize difficulties for a generic model of the real data distributions in their entire frequency range [8], and often do not consider the complete frequency distributions, e.g. an analysis of low frequency words is often omitted. These difficulties reinforce the importance of empirical approaches, leading to many models ([1,2,5,6,10], among others as surveyed in [7]). Still, due to its simplicity and universality, Zipf's law is widely used, as a first approximation or as a basis for improvements. There is a lack of studies (e.g. [4]) on n-grams ($n > 1$), carrying in their specifics a more focused semantic content, useful for relevant multiword extraction. Also, most studies are limited to *corpora* below a few million words. Due to the large orders of magnitude of the language vocabularies, much larger *corpora* are needed to investigate the n-gram behavior for $n \geq 1$. There are recent studies on large *corpora*, [3,9] but often exclude the lower n-gram frequencies, e.g., below 40, as in the 1.8 Tera word Google English n-gram *corpus* [3] for $1 \leq n \leq 5$, precluding model validation with real data.

3 Estimating the Number of Distinct *n*-grams

We assume that the size of each language n-gram vocabulary, e.g., English, is practically fixed at each given temporal epoch, as new/old n-grams slowly emerge/disappear. For brevity, we omit an indication of the language L (English, French) and the n-gram size n (1..6) in the expressions but always assume each expression holds for a given (L, n) pair. Let $V(L, n)$ (denoted as V) be the language n-gram vocabulary size for each n-gram size ($1 \leq n \leq 6$); and $D(C; L, n)$ (denoted as D) be the number of distinct n-grams in a *corpus* of size C in language L, for each given n. We propose a model for estimating D that, as in *growth and preferential attachment models* [5,10], considers two processes: i) selecting new words from a language vocabulary; ii) or repeating existing words in a *corpus*. The model makes it explicit how the vocabulary finiteness influences the rate of appearance of new distinct n-grams as the *corpus* size grows. Regarding i) we follow [5] (whose complete model only applies to character-formed languages, e.g. Chinese) in the particular way those authors model the distinct n-grams from the language vocabulary that are still to appear in the *corpus*, represented by $F_1 = (V - D)/V$. Ratio F_1 monotonically decreases when the *corpus* grows, as the number of distincts (D) approaches the vocabulary size (V). Regarding ii), ratio $F_2 = C/D$ is the average number of occurrences per distinct n-gram. The larger F_2, the stronger the tendency is for repeating existing n-grams in the *corpus*. Thus, we propose the rate of appearance of new distinct n-grams for each (L, n) to be $\propto F_1 \times 1/F_2$, that is the outcome of multiplying F_1 by the reciprocal of F_2. Assuming the validity of a continuum approximation, this rate corresponds to $\frac{\mathrm{d}D}{\mathrm{d}C}$. Let $K_1 > 0$ be a real constant,

$$\frac{\mathrm{d}D}{\mathrm{d}C} = K_1 \, \frac{D}{C} \, \frac{V - D}{V} \quad \Rightarrow \quad \frac{V}{K_1 \, (V - D) \, D} \, \mathrm{d}D = \frac{1}{C} \, \mathrm{d}C \quad \Rightarrow$$

$$\int \frac{V}{K_1 \, (V - D) \, D} \, \mathrm{d}D = \int \frac{1}{C} \, \mathrm{d}C \quad \Rightarrow \quad -\frac{\ln(|\frac{V}{D} - 1|)}{K_1} + c_1 = \ln(|C|) + c_2$$

with c_1, c_2 as integration constants. As $|\frac{V}{D}| \geq 1$ and $C > 0$, with $c_2 - c_1 = \ln(K_2)$,

$$\ln((\frac{V}{D} - 1)^{-\frac{1}{K_1}}) = \ln(K_2) + \ln(C) \quad \Rightarrow \quad (\frac{V}{D} - 1)^{-\frac{1}{K_1}} = K_2 \, C$$

Thus, the number of distinct n-grams for each (L, n) is

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

In Sect. 4, V, K_1 and K_2 are empirically determined for each (L, n) pair.

3.1 Reviewing Zipf's Law

By Zipf's Law [11], the frequency of the r^{th} most frequent word in a *corpus* is

$$f(r) = f(1) \times r^{-\alpha} \tag{2}$$

where r is the word rank (from 1 to $D(L, 1)$) and α is a constant close to 1.

Observations in a wide range of large *corpora* show that the relative frequency of the most frequent 1-gram in English, "the", has small fluctuations around 0.06, being a fair approximation to its occurrence probability, p_1. Thus, $f(1) \approx p_1 C$. From (2), $\ln(f(r)) = \ln(f(1)) - \alpha \ln(r)$ so, ideally, $\ln(f(r))$ decreases linearly with a slope α, as $\ln(r)$ increases. However, in general, real data show deviations from a straight line (e.g. Fig. 1 for real *corpora*: 62; 508; 8 600; in millions of words).

The *steps* in the higher ranks in Fig. 1 correspond to equal-frequency words forming frequency levels (groups) of integer frequency k and size $W(k)$. Only for the lowest k values starting from 1, there are frequency levels with multiple ranks. Figure 2 shows the log-log curve $W(k)$ *versus* k ($k \geq 1$), which can be approximated by a power law, but also exhibiting deviations from real data in both extremes of k [2,6].

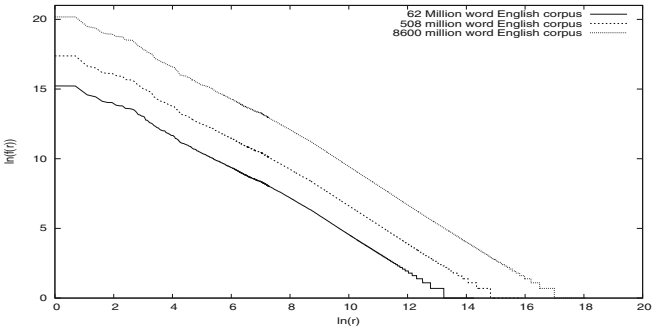

Fig. 1. The observed word rank-frequency distributions

3.2 Estimating the Size $W(k)$ of Each Frequency Level

Considering a generic level k, let r_{l_k} and r_{h_k} be its lowest and highest ranks. Thus, $f(r_{l_k}) = f(r_{h_k}) = k$. This model for estimating $W(k)$ only applies to the higher ranks region of the real data distribution, as long as adjacent frequency levels have consecutive integer frequency values, that is $f(r_{h_{k+1}}) = f(r_{h_k}) + 1$. We follow the functional structure of (2) due to its simplicity and assume, based on empirical observations, that it applies to n-grams $n \geq 1$, with α dependent on n for each language L, although we omit this in the expressions. In a first step, we assume an *ideal* straight line Zipf plot with slope α_z. In further steps, to address the Zipf plot deviations we model the dependencies of the α parameter on the *corpus* size and the level frequency. Thus: $f(r_{h_{k+1}}) = f(r_{h_k}) + 1 = f(1)\, r_{h_{k+1}}^{-\alpha_z} = f(1)\, r_{h_k}^{-\alpha_z} + 1$, leading to

$$r_{h_{k+1}} = \left(\frac{f(1)\, r_{h_k}^{-\alpha_z} + 1}{f(1)} \right)^{-\frac{1}{\alpha_z}} = \left(\frac{1}{r_{h_k}^{\alpha_z}} + \frac{1}{f(1)} \right)^{-\frac{1}{\alpha_z}} . \tag{3}$$

By analogy

$$r_{h_{k+2}} = \left(\frac{1}{r_{h_{k+1}}^{\alpha_z}} + \frac{1}{f(1)}\right)^{-\frac{1}{\alpha_z}} = \left(\frac{1}{\left(\left(\frac{1}{r_{h_k}^{\alpha_z}} + \frac{1}{f(1)}\right)^{-\frac{1}{\alpha_z}}\right)^{\alpha_z}} + \frac{1}{f(1)}\right)^{-\frac{1}{\alpha_z}}.$$

$$r_{h_{k+2}} = \left(\frac{1}{\left(\frac{1}{r_{h_k}^{\alpha_z}} + \frac{1}{f(1)}\right)^{-1}} + \frac{1}{f(1)}\right)^{-\frac{1}{\alpha_z}} = \left(\frac{1}{r_{h_k}^{\alpha_z}} + \frac{2}{f(1)}\right)^{-\frac{1}{\alpha_z}}. \quad (4)$$

$$r_{h_{k+3}} = \left(\frac{1}{\left(\left(\frac{1}{r_{h_k}^{\alpha_z}} + \frac{2}{f(1)}\right)^{-\frac{1}{\alpha_z}}\right)^{\alpha_z}} + \frac{1}{f(1)}\right)^{-\frac{1}{\alpha_z}} = \left(\frac{1}{r_{h_k}^{\alpha_z}} + \frac{3}{f(1)}\right)^{-\frac{1}{\alpha_z}}. \quad (5)$$

So, we can generalize (4) and (5), leading to:

$$r_{h_{k+m}} = \left(\frac{1}{r_{h_k}^{\alpha_z}} + \frac{m}{f(1)}\right)^{-\frac{1}{\alpha_z}}, \quad (6)$$

where $(k+m) : 1..k_{max}$ with $k_{max} = f(1)$, and m is an integer offset starting from 0. With $k = 1$ in (6) we estimate the highest rank of level $k+m$ for each m, as a function of rank r_{h_1}, which is the number of distinct n-grams of size n in the *corpus* ($D(C; L, n)$). So, by subtracting r_{h_k} from $r_{h_{k+1}}$, we estimate for each (L, n), the size $W_z(k)$ (subscript z denotes the Zipf assumption).

$$W_z(k) = \left(\frac{1}{D^{\alpha_z}} + \frac{k-1}{f(1)}\right)^{-\frac{1}{\alpha_z}} - \left(\frac{1}{D^{\alpha_z}} + \frac{k}{f(1)}\right)^{-\frac{1}{\alpha_z}}. \quad (7)$$

To estimate $W_z(k)$, we first calculate $D(C; L, n)$, requiring the K_1 and K_2 constants in (1), Sect. 4.1. Then we tune the α_z value that best fits the $W_z(1)$ value (by (7)) to the observed size $W_{obs}(k)$ of level $k = 1$ in a 508 million word *corpus*. Figure 2 shows the log-log curves of the *Observed word frequency level sizes* for different k for this *corpus* and the $W_z(k)$ *estimates* by (7). The $W_{obs}(k)$ curve exhibits a regular decrease as k grows from 1 until the curve reaches a fluctuation zone, which becomes stronger for higher values of k (discussed ahead in Sect. 4.2). Before the fluctuation zone, the following dominant pattern is suggested: the deviation between the two curves is approximately proportional to $\ln(k)$. This leads us to an improved approximation, denoted by $W(k)$, such that $\ln(W(k)) = \ln(W_z(k)) + \beta \ln(k)$, where β is a real positive constant. Therefore

$$W(k) = W_z(k)\, k^{\beta}. \quad (8)$$

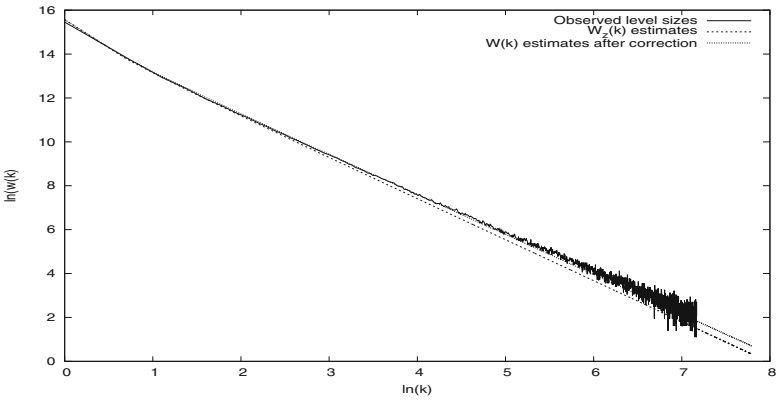

Fig. 2. Word frequency level size $W(k)$ vs k: observed and estimates by (7) and (8).

β is tuned, keeping the *corpus* fixed, to the best fit of $W(k)$ to the observed level sizes. Curve $W(k)$ *estimates after correction* by (8) is much closer to the observed one (Fig. 2, Table 5). Similar behaviors were found for n-grams $1 < n \leq 6$.

3.3 The Effect of the *Corpus* Size on the Level Size $W(k)$

Unlike the constant Zipf's α_z in (7), for real *corpora*, exponent α in (2) depends both on the individual ranks of the distinct n-grams for each *corpus* and on the *corpus* size. Thus, $W_z(k)$ calculation in (7) should cope with the α variation. Previously, the α_z value was tuned to fit $W_z(1)$, the size of level 1, in one of the available *corpora* (e.g. 508 million words *corpus*). This is a practical way to fit, with good approximation to the other frequency levels. The following are (*corpus* size; α_z) pairs obtained, for 1-grams, for a set of different *corpora*: (128 364 577; 1.13848), (254 801 364; 1.14671), (508 571 317; 1.155), (1 068 282 476; 1.16391), (2 155 599 290; 1.17233), (4 278 548 582; 1.18056). These values show that α_z grows approximately an equal amount as the *corpus* size is doubled, suggesting a logarithmic proportionality between α_z and the *corpus* size. Let $Q = \log_2(C_2/C_1)$ and $A = \alpha_2 - \alpha_1$ with α_1, α_2 associated, respectively, to C_1, C_2. Thus $\lambda\, \Delta Q = \Delta A$, where λ is a constant, so $\frac{\mathrm{d}A}{\mathrm{d}Q} = \lambda$ and

$$\int \mathrm{d}A = \lambda \int \mathrm{d}Q \quad \Rightarrow \quad A + ct_1 = \lambda\, Q + ct_2 \quad \Rightarrow \quad A = \lambda\, Q + ct_3 \quad \Rightarrow$$

$$\alpha_2 - \alpha_1 = \lambda\, \log_2(\frac{C_2}{C_1}) + ct_3 \quad \Rightarrow \quad \alpha_2 = \alpha_1 + \gamma\, \ln(\frac{C_2}{C_1}) + ct_3 \qquad (9)$$

where $\gamma = \lambda / \ln(2)$ and $ct3 = ct2 - ct1$. ($ct_3 = 0$ in the experiments.) This leads to (10), with $\alpha_1 = \alpha_z$ for some reference *corpus* with size C_1. Any of the above (*corpus* size; α_z) pairs can be used for this, e.g. for 1-grams in English, the C_1 and α_1 values of 508 571 317 and 1.155.

$$\alpha(C) = \alpha_1 + \gamma \ln(\frac{C}{C_1}). \tag{10}$$

$\alpha(C)$ replaces α_z for calculating $W_z(k)$ in (7), now changing to $W_z(k, C)$:

$$W_z(k, C) = \left(\frac{1}{D^{\alpha(C)}} + \frac{k-1}{f(1)}\right)^{-\frac{1}{\alpha(C)}} - \left(\frac{1}{D^{\alpha(C)}} + \frac{k}{f(1)}\right)^{-\frac{1}{\alpha(C)}}, \tag{11}$$

which is reflected in $W(k, C)$ of (8):

$$W(k, C) = W_z(k, C)\, k^{\beta}. \tag{12}$$

Equation (12) allows to predict the k-level size for n-grams $1 \leq n \leq 6$, given C. All expressions (1)–(12) apply to n-grams $1 \leq n \leq 6$ for *corpora* in a language L. The obtained α_z values are lower as the n-gram size increases from 1 to 6.

4 Results

The *corpora* were built by random extraction of English and French Wikipedia documents. For English, the *corpora* sizes were doubled successively, from 62×10^6 words (62 Mw) to 8.6×10^9 words (8.6 Gw). For French, from 71 Mw to 2.4 Gw. Exact English *corpora* sizes are 62 557 077; 128 364 577; 254 801 364; 508 571 317; 1 068 282 476; 2 155 599 290; 4 278 548 582; 8 600 180 252; denoted, respectively, as 62 Mw; 1/8 Gw; 1/4 Gw; 1/2 Gw; 1.1 Gw; 2.2 Gw; 4.3 Gw; 8.6 Gw. French sizes are 71 083 803; 142 889 828; 289 392 085; 595 034 875; 1 154 477 213; 2 403 390 530. Due to lack of space, only English results are shown, French ones being similar.

For a fair count of the distinct n-grams, still not modifying the text semantics, the *corpora* were pre-processed by separating words, through a space, from each of the following characters: {'<', '>', '"', '!', '?', ':', ';', ',', '(', ')', '[', ']'}. Table 1 shows the *corpora* sizes and the distinct n-gram counts.

Table 1. The observed number of distinct n-grams for each *corpus* in English

Corpus	1-grams	2-grams	3-grams	4-grams	5-grams	6-grams
62 Mw	1 567 905	7 682 911	17 507 174	25 872 310	31 427 945	34 998 302
1/8 Gw	2 646 714	12 608 307	29 721 342	45 508 866	56 598 262	64 020 127
1/4 Gw	4 727 634	23 007 130	56 493 059	89 290 502	113 192 655	129 410 481
1/2 Gw	7 905 576	39 045 477	100 093 384	164 049 701	212 902 754	246 756 201
1.1 Gw	15 033 759	74 361 922	199 655 660	341 316 872	454 904 356	534 494 709
2.2 Gw	24 865 840	122 366 976	337 988 348	598 619 341	819 532 204	978 561 656
4.3 Gw	42 363 831	210 708 582	604 996 078	1 113 522 090	1 567 962 556	1 901 784 002
8.6 Gw	70 227 712	350 076 300	1 036 027 979	1 978 136 904	2 866 649 212	3 539 349 002

4.1 Predicting the Number of Distinct n-grams

In order to find K_1, K_2 and $V(L,n)$ for a language L and an n-gram size, to obtain estimates by (1) we start by setting $V(L,n)$ to 10^6 and successively increase it until the K_1 and K_2 values lead to the smallest relative error for two *corpora* of sizes close to the extremes of the *corpora* sizes range: 1/4 G and 4.3 G for English. Relative error is $((Est - Obs)/Obs) \times 100\%$, for estimated ($Est$) and observed ($Obs$) numbers. This procedure stops when further increases of $V(L,n)$ do not lead to significant changes in the relative error, and then that $V(L,n)$ value is taken as an approximation to the n-gram vocabulary size. Table 2 shows K_1 and K_2, and $V(L,n)$ for English. In Table 3, the left sub-column of each n-gram column shows acceptable values for the relative errors of the estimates $D(C;L,n)$ by (1), in this range of *corpora* sizes. The right sub-column shows the estimates of a Poisson based model given by $Dist_{Poisson}(L,n,C) = \sum_{r=1}^{r=V(L,n)}(1 - e^{-f(r,L,C))})$; r is the n-gram rank and $f(r,L,C))$ is the expected frequency of r in corpus of size C by Zipf-Mandelbrot model (see [9], where the parameters were tuned by the same procedure as described above in this section, for the empirical model). For n-gram sizes lower than 4, relative errors are considerably higher than the ones by $D(C;L,n)$, e.g. reaching -31.1%, -24.5% and -12.4% for the largest *corpus* (8.6 Gw). For 5-grams and 6-grams, relative errors are lower. Figure 3 shows that the curves for

Table 2. K_1, K_2 and vocabulary sizes (V, in number of n-grams) for English

	1-grams	2-grams	3-grams	4-grams	5-grams	6-grams
K_1	0.8377775	0.8607456	0.8845	0.9235369	0.9376907	0.955206
K_2	3.61×10^{-11}	5.1×10^{-11}	2.66×10^{-11}	1.7835×10^{-11}	4.29×10^{-12}	6.5×10^{-13}
V	2.45×10^{8}	9.9×10^{8}	4.74×10^{9}	1.31×10^{10}	6.83×10^{10}	5.292×10^{11}

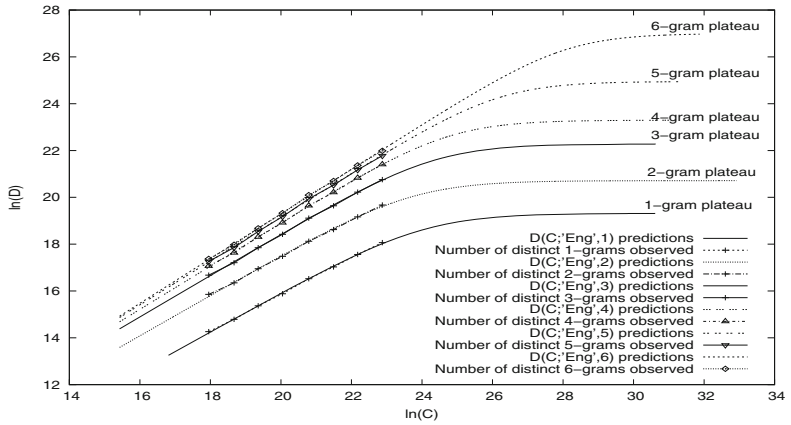

Fig. 3. Numbers of distinct n-grams: observed and predicted ($D(C;L,n)$, by (1)), *versus* the *corpus* size, in English.

Table 3. Relative errors (%) of English distinct n-grams, estimated by: $D(C; L, n)$, (1) (bold left col.); $Dist_{Poisson}(L, n, C)$, [9] (right col.)

Corpus	1-grams		2-grams		3-grams		4-grams		5-grams		6-grams	
62 Mw	**−5.8**	9.2	**−9.1**	−2.0	**−6.0**	−7.2	**5.1**	−9.4	**−2.7**	−5.3	**−3.3**	−5.6
1/8 Gw	**1.5**	7.8	**2.2**	4.8	**4.3**	3.0	**4.5**	4.1	**6.0**	4.8	**5.0**	4.4
1/4 Gw	**0.0**	−2.1	**−3.8**	−3.3	**0.0**	4.1	**0.0**	−3.0	**0.7**	0.0	**0.0**	0.0
1/2 Gw	**5.1**	−5.4	**4.8**	−4.2	**3.0**	0.0	**2.4**	0.0	**2.2**	−2.3	**1.5**	1.5
1.1 Gw	**0.0**	−17.3	**1.0**	−13.3	**−2.4**	−3.2	**−3.5**	−7.3	**−4.4**	−4.5	**−4.9**	−4.5
2.2 Gw	**3.8**	−20.1	**5.2**	−13.2	**3.6**	−7.9	**2.9**	−2.5	**1.9**	0.0	**1.5**	1.3
4.3 Gw	**0.0**	−26.8	**−0.2**	−19.7	**0.0**	−10.2	**0.0**	−6.6	**0.2**	−2.3	**0.4**	0.0
8.6 Gw	**−4.8**	−31.1	**−6.8**	−24.5	**−2.4**	−12.4	**−0.4**	−7.8	**3.3**	−1.3	**4.7**	2.2

the observed and estimated values are quite close, for each n-gram size, across the analysed *corpora*. The predictions extend beyond the empirical *corpora* range, evolving to the n-gram vocabulary sizes *plateaux*. Equation (1) predicts, e.g., about 99% of the distinct n-grams in each English n-gram vocabulary appear for $C \approx 6.7 \times 10^{12}$ words for 1-grams, and $C \approx 1.89 \times 10^{14}$ words for 6-grams.

4.2 The Frequency Level Sizes

Table 4 shows the β and γ values for calculating $W(k, C)$ (12) and $\alpha(C)$ (10), $1 \leq n \leq 6$. Equation (12) provides a good approximation when the observed level sizes, $W_{obs}(k, C)$, decrease monotonically as k grows: $W_{obs}(k, C) > W_{obs}(k + 1, C)$. For a fixed *corpus* size, that condition is not ensured when k exceeds a certain threshold, which is lower for smaller *corpora* and also for smaller n-gram sizes. E.g., for the 62 Mw *corpus* the k threshold is 28 for 1-grams, and is 145 in the 8.6 Gw *corpus* for 2-grams. Above k threshold, due to $W_{obs}(k, C)$ non-monotonicity, model (12) only provides a rough approximation (Fig. 4), in contrast to its good approximation in lower k. Table 5 shows error metrics for the $W(k, C)$ estimates, considering the following basic set of k values: $k \in K$, $K = \{1, 2, 3, 4, 5, 6, 7, 8, 16, 32, 64, 128\}$. Due to the k thresholds, the full set of k values was only used for the two corpora whose denoted sizes are above 4 Gw; the k value of 128 was not considered for the 1.1 Gw and 2.2 Gw *corpora*; 128 and 64 were not used for the 1/2 Gw and 1/4 Gw ones; 128, 64 and 32 were not considered for the remaining *corpora*, 1/8 Gw and 62 Mw. Table 5 shows two columns for each n-gram size: the left one indicates the average relative error,

Table 4. Parameters β and γ for each n-gram size and English

	1-grams	2-grams	3-grams	4-grams	5-grams	6-grams
$\boldsymbol{\beta}$	0.06	0.874	0.11	0.167	0.15	0.128
$\boldsymbol{\gamma}$	0.011	0.012	0.0115	0.012	0.01205	0.01205

Table 5. Error metrics (percentages): *i*) average relative error (absolute value) of $W(k, C)$ estimates for English; *ii*) *root-mean-squared-deviation* of the relative error.

Corpus	1-grams		2-grams		3-grams		4-grams		5-grams		6-grams	
62 Mw	5.5	6.1	2.7	3.5	5.9	6.6	7.9	8.9	4.2	5.9	8.2	9.4
1/8 Gw	7.9	8.3	7.8	8.1	3.1	4.0	3.6	5.0	4.1	4.9	7.5	8.0
1/4 Gw	6.1	6.4	5.4	6.0	2.5	3.0	5.2	5.7	4.7	5.4	3.0	4.6
1/2 Gw	6.8	7.2	8.3	8.9	4.7	5.4	6.6	6.8	6.4	6.7	5.4	6.2
1.1 Gw	2.4	3.2	6.1	6.9	3.8	4.4	3.4	4.0	4.0	4.7	4.1	5.5
2.2 Gw	3.4	4.3	7.7	8.4	7.7	8.5	8.3	8.7	8.1	9.2	4.6	5.3
4.3 Gw	3.4	4.8	6.1	6.9	6.9	7.8	5.9	6.1	6.5	7.4	4.1	6.8
8.6 Gw	6.2	7.4	5.6	6.1	7.2	8.5	6.0	6.7	6.1	7.5	5.8	8.8

$\frac{1}{\|K\|}\sum_{k\in K} Err(k)$, where K is the set of k values used in the *corpus* as explained before, and $Err(k) = |(W(k, C) - W_{obs}(k, C))/W_{obs}(k, C)|$, which is the relative error (in its absolute value) for each k; each value in the right column, based on the *root-mean-squared-deviation*, is calculated as $\sqrt{\frac{1}{\|K\|}\sum_{k\in K} Err(k)^2}$ and reflects how homogeneous the values of the relative error are for the different k values used in the estimates. The closer the left and right column values are, the greater the homogeneity. Global results exhibit homogeneity, showing also reasonably low average values. It should also be noted that the considered range of k values includes in all cases the lower frequency values, at least from 1 to 16.

Figure 4 shows, e.g., for 1-grams and 3-grams, that the curves $W(k, C)$ ("Estimated") are very close to the curves $W_{obs}(k, C)$ ("Observed") for each *corpus*. Likewise for other n-gram sizes (2, 4, 5, 6). Beyond the k thresholds the observed curves $W_{obs}(k, C)$ enter non-monotonic fluctuation zones. The slope of the curves changes only slightly as the corpus size grows and the similar spacing between curves when C is doubled reflects a regular $W_{obs}(k, C)$ growth pattern.

4.3 The Evolution of the Frequency Level Sizes

Model (12) predicts that, for each of the lowest k values, $W(k, C)$ grows with C until a maximum, then gradually decreases with increasing C (Fig. 5). This results from the language vocabulary finiteness. E.g., for the singletons, $W(1, C)$ keeps growing with C while n-grams remain to appear from the vocabulary. For a large enough *corpus*, the distinct n-gram *plateau* is reached (Fig. 3) and after this point, $W(1, C)$ can not grow anymore. By further increasing C, the existing singleton n-grams will gradually move to the frequency level $k = 2$, until $W(1, C) = 0$, i.e. singletons disappear. Similar behavior is predicted for $1 < n \leq 6$ (Fig. 5 a). By further increasing C, this process will affect successive k levels, e.g. $k = 2$ and so on, e.g., Fig. 5 b).

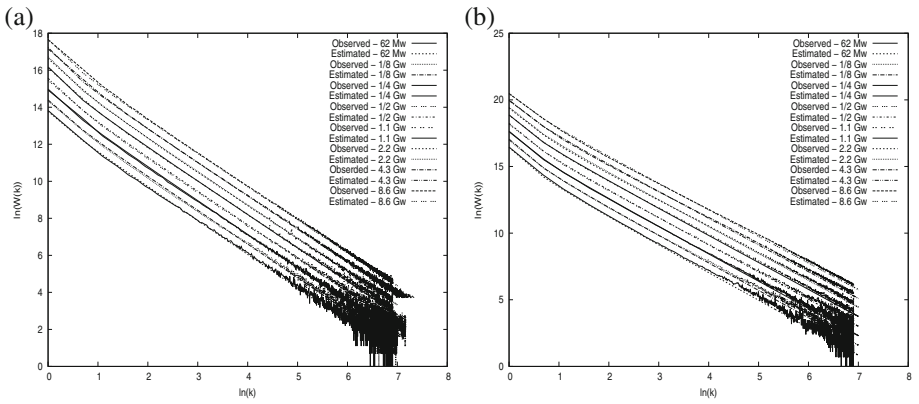

Fig. 4. Observed ($W_{obs}(k, C)$) and estimated ($W(k, C)$) (by (12)) values for different frequency levels and *corpora* sizes—English: a) 1-grams; b) 3-grams.

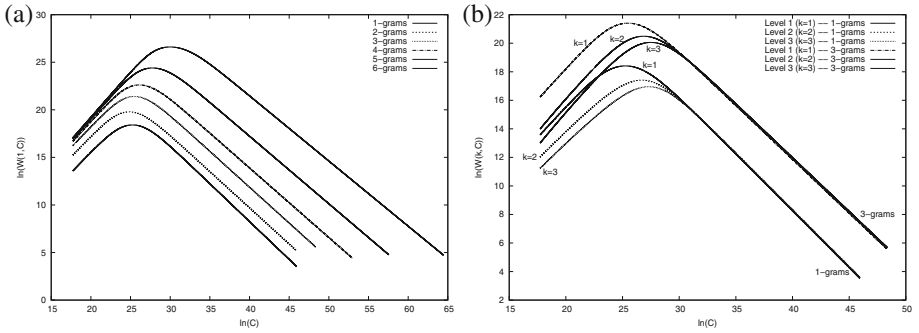

Fig. 5. $W(k, C)$ vs C: a) $k = 1$, 1-gram..6-gram; b) $k \in \{1, 2, 3\}$, 1-gram, 3-gram.

5 Conclusions

Statistical extraction of relevant multiwords benefits with n-gram frequency distribution information from real *corpora*. This goes beyond the usual word frequency distribution (i.e. 1-grams) by including n-grams of sizes $n > 1$. This paper contributes with an empirical study on the n-gram frequency distribution from 1-grams to 6-grams, with large real *corpora* (English and French) from millions up to a few billion words. A distinctive aspect is that it analyzes the low frequency n-grams real data for such large *corpora* instead of relying on smoothing-based estimation. Low frequency n-grams represents the largest proportion of the distinct n-grams in a corpus for a wide range of *corpora* sizes, and are a significant part of the most relevant multiwords. This paper contributes with an empirical analysis and a model of the properties of the low frequency n-grams in large corpora, complementing studies on low frequency single words for smaller *corpora*. Assuming the finiteness of language n-gram vocabularies,

we analyzed and modelled the total number of distinct n-grams for the above range of *corpora*. The model leads to good approximations to the real data distributions, with average relative errors of 5.6% and 2.9% respectively for the lower frequency n-gram distribution (namely the number of singleton n-grams), and the number of distinct n-grams. Moreover, the proposed model allows to predict the evolution of the numbers of distinct n-grams towards asymptotic *plateaux* for large enough *corpora*. Also, according to the model, the sizes of equal-frequency levels, for the lowest frequencies, initially grow with the *corpus* size until reaching a maximum and then decrease as the *corpus* grows to very large sizes. Overall, these results have practical implications for the estimation of the capacity of n-gram Big Data systems. Work is ongoing towards extending this empirical study up to hundreds of Giga word *corpora*.

References

1. Ausloos, M., Cerqueti, R.: A universal rank-size law. PLoS ONE **11**(11) (2016). https://doi.org/10.1371/journal.pone.0166011
2. Balasubrahmanyan, V.K., Naranan, S.: Algorithmic information, complexity and Zipf's law. Glottometrics **4**, 1–26 (2002)
3. Brants, T., Popat, A.C., Xu, P., Och, F.J., Dean, J.: Large language models in machine translation. In: Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning, pp. 858–867. ACL (2007)
4. Dias, G.: Multiword unit hybrid extraction. In: ACL Workshop on Multiword Expressions, vol. 18, pp. 41–48. ACL (2003). https://doi.org/10.3115/1119282.1119288
5. Lü, L., Zhang, Z.K., Zhou, T.: Deviation of Zipf's and heaps' laws in human languages with limited dictionary sizes. Sci. Rep. **3**(1082) (2013). https://doi.org/10.1038/srep01082
6. Mandelbrot, B.: On the theory of word frequencies and on related Markovian models of discourse. In: Structure of Language and its Mathematical Aspects (1953)
7. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. Internet Math. **1**(2), 226–251 (2003)
8. Piantadosi, S.T.: Zipf's word frequency law in natural language: a critical review and future directions. Psychon. Bull. Rev. **21**(5), 1112–1130 (2014). https://doi.org/10.3758/s13423-014-0585-6
9. Silva, J.F., Gonçalves, C., Cunha, J.C.: A theoretical model for n-gram distribution in big data corpora. In: 2016 IEEE International Conference on Big Data, pp. 134–141 (2016). https://doi.org/10.1109/BigData.2016.7840598
10. Simon, H.: On a class of skew distribution functions. Biometrika **42**(3/4), 425–440 (1955). https://doi.org/10.2307/2333389
11. Zipf, G.K.: Human Behavior and the Principle of Least-Effort. Addison-Wesley, Cambridge (1949)