A Scalable Model for Frequency Distribution of Low Occurrence Multi-words Towards Handling Very Large Spectrum of Text *Corpora* Sizes

Joaquim F. Silva (✉)[0000−0002−5223−1180] and Jose C. Cunha

NOVA LINCS, NOVA School of Science and Technology, email{jfs,jcc}@fct.unl.pt

Abstract. Predicting the diversity of words and multi-words (n-grams) in a text *corpus* and their frequency distributions is important in NLP and language modeling, and is becoming critical to enable the design of modern applications, namely Large Language Models, e.g. for guiding tokenization and *corpus* analysis for pre-training. This requires the ability to model the very large scale *corpora* behaviour, the handling of multi-words as subwords or phrases, and the distribution of n-grams across different frequency ranges, namely the low occurrence n-grams.
We present a scalable model to predict the number of distinct n-grams and their frequency distributions targeting an extended range of *corpora* sizes, from hundreds of million words to hundreds of billion words (a 1000 times factor). This led us to a novel approach for explicitly incorporating into the model the parameter dependency behaviour regarding the extended *corpora* size range.
In the presence of such extended range of *corpora* sizes, the model estimates the cumulative numbers of distinct n-grams ($1 \leq n \leq 6$) greater or equal to a given frequency $k \geq 1$, in a *corpus*, and the numbers of n-grams with equal-frequencies, in a given language *corpus*. Unlike most approaches that assume an open, potentially infinite, language word vocabulary, this model relies on the vocabulary finiteness. The model ensures very low and stable average relative errors (*circa* 2%), for the low frequencies starting with singletons, from 1-grams to 6-grams, across the above very large range of *corpora* sizes, in English and German.

Keywords: Scalable Prediction Model · Large Text Corpora · n-gram Frequency Distribution.

1 Introduction

A word n-gram is a sequence of n consecutive words. Knowledge on the statistical n-gram frequency distributions in text *corpora* is useful in applications, e.g. indexing, extracting relevant terms, compression, cache design, and translation. In large language modelling, understanding n-gram distributions as a function of *corpus* size is useful for: *a*) Guiding Tokenization Strategies, as tokenizers can be tuned to produce subwords or phrases that approximate frequent n-gram patterns; besides, the balance of distinct 1-grams (words) *versus* higher

order n-grams (multi-words) in different frequency ranges can help determine the appropriate tokenization granularity; b) *Corpus* Analysis for pre-training, where knowing the number of distinct n-grams across frequency ranges serves as a proxy for a *corpus* linguistic diversity; a *corpus* with n-grams spanning a broad range of frequencies likely captures richer patterns, important for pre-training deep language models; however, if the numbers of distinct n-grams stabilise beyond a certain *corpus* size, it suggests diminishing returns when adding more data, guiding the selection of an optimal *corpus* size for pre-training or fine-tuning.

Most traditional frequency distribution models consider moderate size *corpora* (from thousands to several million words (Mw)) and only apply to single words (1-grams). However, larger *corpora* have impact upon n-gram frequency distributions, as shown by the emergence of Big Data. Also, multi-word n-grams ($n \geq 2$) reveal the language phrase/subphrase structure and express semantic specificity, and are becoming more relevant in an increasingly number of applications. Furthermore, most of the semantic content words and multi-words, e.g. important for topic mining, appear in the low frequency range, occurring 1,2,3... times, and they represent the majority of the distinct n-grams in each given *corpus*. In [11] a language-independent model is proposed for words and multi-words (from 1-grams to 6-grams) of low occurrence frequencies. It predicts the cumulative number of distinct n-grams, $D(k;C)$, with frequencies greater than or equal to k, for $k \geq 1$, in a *corpus* of size C, ($D(C) = D(1,C)$ is the total number of distinct n-grams), as well as the sizes, $W(k,C)$, of groups of n-grams with equal frequencies, as a function of *corpus* size. The principles underlying the model have a great potential for applications mainly when considering extremely large *corpora* sizes, handling words and multi-words, and low-frequency n-grams. All the above motivates the overall goal of this paper, that is to further explore the rationale behind the above model – that we denote as the baseline model – and evaluate its adequacy to predict $D(C)$, $D(k,C)$ and $W(k,C)$ variables, in a wider range (spanning a 1×1000 factor) of very large *corpora*, namely going into the hundred billion words (Gw) scale. The main contributions of this paper are the achievement of very low and stable average relative errors (around 2%) in the prediction of the above variables, encompassing 1-grams to 6-grams, and for the low frequencies starting with singletons, across the above very large range of *corpora* sizes, in English and German. This was achieved by considering the dependence of the model parameters on *corpus* size and their fine tuning, enforcing a sound estimation methodology relying on the separations of the training/validation and testing *corpora*, and by proposing a well-founded method for identifying the frequency limits of the model validity. This is in contrast to the baseline approach, which exhibits a critical issue when assuming the constancy of the model parameters versus C, and whose usage for *corpora* well beyond the 8.6 Gw largest *corpus* size in [11], revealed its inadequacy for large scale *corpora*, having led to significant deviations from real data. Besides, the usage of the same *corpora* sets in the baseline, both for training and testing purposes, is inadequate. We present the background of this work, the new proposed approach,

experimental results and conclusions. A guide for the model reproducibility is found at https://github.com/OurName1234/ngrams .

2 Background and Related Work

Several influential word frequency distribution models were proposed [9], including the empirical Zipf's Law [13], showing the word frequency distribution as an approximated power law, but deviating from real data for high and low frequencies, and also theoretical models, e.g. based on preferential attachment [12]. However, firstly, most models only consider word frequencies, ignoring multi-words, although language modeling benefits from the knowledge on n-gram frequency patterns [6, 4, 5, 10, 11]. Secondly, the model predictions often show deviations from the real *corpora* data, in the high and low occurrence frequencies. Often, they ignore the low-frequency words, or are unable to accurately model the large set of less frequent, content words in a *corpus*, being important in many applications. Thirdly, most models have been tested only with small and moderate size *corpora* (up to several million words). However, the emergence of BigData and Web-based very large *corpora* and/or n-gram frequency data [2, 3] triggered the development of large-scale applications [2], posing new challenges.

We address a challenge posed by large text *corpora* concerning the growth of the available *corpora* sizes and their effect upon the numbers of distinct n-grams and their frequency distributions, for evaluating the models/applications. The evolution of the number of distinct words (D) wrt to the *corpus* size (C), is modeled by Herdan's and Heaps'empirical law [1], assuming an infinite word language vocabulary and stating that D would always keeps growing with increasing *corpus* sizes, as a power law with a constant exponent, but empirical evidence from large *corpora* shows that such exponent depends on the *corpus* size [1], suggesting that D will eventually saturate as C tends to infinity. This saturation of D occurs in word frequency distributions in languages with limited word vocabularies, eg. Chinese, Japanese, Korean [8]. There is a lack of models predicting how the *corpus* size, for a wide range of large *corpora* sizes, explicitly influences the $D(C)$, $W(k, C)$ and $D(k, C)$ distributions for multi-words, namely considering the low occurrence n-grams and the model validation with real large *corpora* from different languages. This is useful to predict the impact of *corpus* growth upon application time and space complexities, thus supporting application design. Only a few models [5, 10, 11] address the above issues by unified approaches. However, only [11] relies on a principled model – the baseline model –, reflecting to the best of our knowledge, the state of the art of unified approaches for predicting the effect of *corpus* size upon the n-gram frequency distributions, for low frequencies and a wide range of large *corpora* sizes.

3 The Proposed Approach

Brief Review of the Baseline Model. The baseline [11] model assumes that for a fixed temporal epoch, there is an n-gram language L vocabulary

with size $V(L,n)$. $D(k,C;L,n)$ is the number of distinct n-grams of size n, occurring at least k times in a *corpus* of size C of language L, also denoted $D(k,C)$ when L and n are implicit. For $k=1$, $D(k,C;L,n)$ is denoted as $D(C;L,n)$ or $D(C)$ if L and n are implicit. Under a continuum approximation, the growth rate of $D(k,C;L,n)$ wrt C, with $k \geq 1$, is modeled by the derivative $dD(k,C;L,n)/dC$, influenced by two factors: one is inspired by a cumulative form of preferential attachment, such that, when the *corpus* size C grows by a given amount, each $D(k,C;L,n)$ tends to increase at a rate proportional to $D(k,C;L,n)/C$, its current relative size in the *corpus*; another is due to the finiteness of the n-gram language vocabulary $V(L,n)$, reflecting a slowdown effect defined by the proportion of remaining n-grams still having a frequency below k, regardless of whether they appear in the current *corpus* or are unseen n-grams: $(V(L,n) - D(k,C;L,n))/V(L,n)$. For $k=1$, this is the proportion of the finite vocabulary n-grams still unseen in the current *corpus* of size C. Thus, $\frac{dD(k,C;L,n)}{dC}$ is given by:

$$\frac{dD(k,C)}{dC} = g_k \frac{D(k,C)}{C} \frac{V - D(k,C)}{V} \tag{1}$$

where V and g_k simplify $V(L,n)$ and the proportionality factor $g_k(L,n)$ respectively. Indeed, $V = \sum_{k=1}^{k=kmax} W(k,C)$ where $W(k,C)$ is the number of distinct n-grams with frequency k and $kmax$ is the highest frequency in the *corpus*, for each n. The solution of equation (1) is (h_k standing for an integration constant):

$$D(k,C;L,n) = \frac{V(L,n)}{1 + (h_k(L,n)\,C)^{-g_k(L,n)}} \ . \tag{2}$$

From (2), $W(k,C;L,n)$, the number of equal-frequency (k) distinct n-grams of size n, is predicted by the subtraction of the cumulative numbers $D(k,C;L,n)$ and $D(k+1,C;L,n)$ for two frequency consecutive values, k and $k+1$.

3.1 An Approach to Large-scale Corpora

The baseline model [11] was trained for English with corpora up to 8.6 Gw and would be able to predict $D(k,C;L,n)$ and $W(k,C;L,n)$ values for any *corpus* size with average relative errors around 3%. However, for the purpose of evaluating that model (as available at "http://bit.ly/3gqM6rS") for extended large *corpora* ranges, we experimented with large scale English *corpora* reaching hundreds of billion words, and the obtained predictions show significant deviations from the empirical values, reflecting relative errors with modules much larger than 3% (generally over 20%). This led us to a new approach considering the dependency of the model parameters $g_k(L,n)$ and $h_k(L,n)$ on the *corpus* size C. Indeed, by considering the above dependencies, we achieved significantly lower errors in the model predictions compared to the baseline. Timewise, this involves typical n-gram counting in large *corpora* – which is computationally heavy, requiring the use of a parallel computing infrastructure – but is done only once, for model parameter estimation, while model utilisation for prediction purposes

only needs a fast formula calculation (2). The errors obtained by the new approach kept stable across a much wider range of large *corpora* sizes, spanning a 1×1000 factor, for English, and reaching 373 billion words – a comparison of the relative errors when assuming the $g_k(L,n)$ and $h_k(L,n)$ constancy [11], *versus* when considering their dependency on C is presented on Sect. 5.

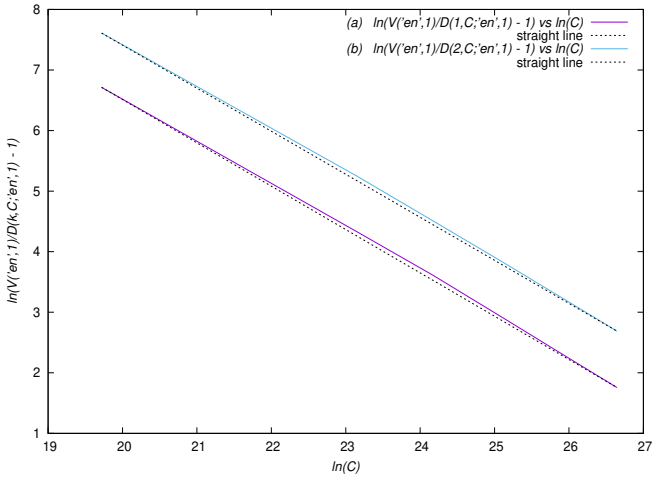


Fig. 1: Dependency of $\ln(\frac{V(L,n)}{D(k,C;L,n)} - 1)$ *versus* $\ln(C)$ for empirical 1-gram counts in English *corpora* (solid lines). Dashed lines refer to constancy assumption.

The Parameters Constancy Assumption in the Baseline Model. From the baseline model (2), the curve for $\ln(\frac{V(L,n)}{D(k,C;L,n)} - 1) = \ln((h_k(L,n)C)^{-g_k(L,n)}) = -g_k(L,n)\ln(h_k(L,n)) - g_k(L,n)\ln(C)$ is a straight line with slope $-g_k(L,n)$ when drawn as function of $\ln(C)$, because the constancy of $g_k(L,n)$ and $h_k(L,n)$ *versus* C is assumed. However, the experimental curves of $\ln(\frac{V(L,n)}{D(k,C;L,n)} - 1)$ *versus* $\ln(C)$, for the empirical $D(k,C;L,n)$ values (for English 1-grams and $k \in \{1,2\}$), show noticeable deviations from straight lines (in Fig. 1 the experimental curves are solid and straight lines are dashed): this visual perception is consistent with the large numeric values of the relative errors obtained when using the baseline model, as reported in Sect. 5, tables 1, 2, 3, 4, 5. Note that this figure covers an extended *corpora* range of test *corpora* (beyond the range considered in the baseline): 366 Mw, 11.3 Gw, 31.5 Gw, 82.7 Gw, 172 Gw and 373 Gw. In this experiment, we considered an estimated English 1-gram vocabulary of $V('en', 1) = 2.95e9$ – as discussed in Sect. 3.2. Thus, we model $D(k,C;L,n)$ and $W(k,C;L,n)$ with the explicit dependency $g_k(C;L,n)$ and $h_k(C;L,n)$:

$$D(k,C;L,n) = \frac{V(L,n)}{1 + (h_k(C;L,n) \cdot C)^{-g_k(C;L,n)}} \quad . \tag{3}$$

$$W(k, C; L, n) = D(k, C; L, n) - D(k+1, C; L, n) \tag{4}$$

3.2 Estimating the Model Parameters

We present the method for estimating the model parameters – for each (L, n) –, considering the dependence on C: $V(L, n)$, $g_k(C; L, n)$ and $h_k(C; L, n)$. We use a set of training *corpora* to estimate the parameters, and a separate set of testing *corpora* to evaluate the results, with *corpora* up to hundreds of billion words. Overall, this method involves estimating: *i*) the vocabulary size $V(L, n)$; *ii*) the parameters $g_k(C; L, n)$ and $h_k(C; L, n)$ with the training *corpora*; *iii*) the dependency behavior of $g_k(C; L, n)$ and $h_k(C; L, n)$ for general test *corpora*. Concerning *i*), we estimate the vocabulary size to ensure the lowest average relative errors of model $D(k, C; L, n)$ (3) for each k. To avoid the computational complexity of an exhaustive search – which would be $O(S^{2N+1})$, S being the number of considered candidate values for each parameter ($V(L, n)$, $g_k(C; L, n)$, $h_k(C; L, n)$) and N the number of training *corpora* –, in choosing the best parameter combination, we first estimate $V(L, n)$. Concerning *ii*), given the estimated $V(L, n)$, we estimate the pairs $(g_k(C; L, n), h_k(C; L, n))$ for each training *corpus* C. Concerning *iii*), given the collection of the above estimated pairs $(g_k(C; L, n), h_k(C; L, n))$, we rely on regression using *splines* [7] (outperforming piecewise linear methods), thus enabling the estimation of values of $g_k(C; L, n)$ and $h_k(C; L, n)$ for any general *corpora*.

i) Estimating the vocabulary size. If we draw a secant line connecting two points (i and j) on one of the curves (a) or (b) of Fig. 1, whose corresponding *corpora* size values are $\ln(C_i)$ and $\ln(C_j)$, then the slope defined by this secant corresponds to a value of $g_k(C; L, n)$ that is valid for both *corpora*. So, there is a $g_k(C; L, n) = g_{k_{i,j}}$ that can fit both *corpora* C_i and C_j. Let h_{k_i} and h_{k_j} be the h_k parameter values corresponding, in (3), respectively, to the *corpora* C_i and C_j. Let V, D_{k_i} and D_{k_j} abbreviate $V(L, n)$, $D(k, C_i; L, n)$ and $D(k, C_j; L, n)$. Under the assumption $g_k(C; L, n) = g_{k_{i,j}}$, we obtain $g_{k_{i,j}}$ from (3), the left-side equation in (5). Also, under the approximation of assuming a common value, $h_{k_{i,j}}$, for $h_{k_i} \approx h_{k_j}$ (discussed below), we obtain the right-side equation in (5).

$$g_{k_{i,j}} = \ln\left(\frac{(V - D_{k_i})\, D_{k_j}}{(V - D_{k_j})\, D_{k_i}}\right) / \ln\left(\frac{h_{k_j}\, C_j}{h_{k_i}\, C_i}\right) \qquad h_{k_{i,j}} = \frac{1}{C_j((V/D_{k_j}) - 1)^{(1/g_{k_{i,j}})}} \tag{5}$$

When using $D(k, C; L, n)$ (3) for several values of k, there must be a $V(L, n)$ value that leads to the model predictions minimizing the average relative errors for the training *corpora* set. We consider a range of k values from 1 to $kmax$ ($kmax$ set to 2^{12}, explained in Sect. 5.2), and pairs of values $(g_k(C; L, n), h_k(C; L, n))$, drawn from two distinct candidates ranges: one for $g_k(C; L, n)$ varying around an initial g_{init} value, and another range for $h_k(C; L, n)$, varying around an initial h_{init}. So, to find the initial values g_{init} and h_{init}, to be used as starting points for the above search, we apply (5) to obtain $g_{init} = g_{k_{i,j}}$ and

$h_{init} = h_{k_{i,j}}$. Thus, let C_{min} and C_{max} be the smallest and the largest *corpora* from the training set. In Eq. (5) we instantiate $C_i = C_{min}$, $C_j = C_{max}$ and the corresponding empirical (D_{emp}) values of $D_{k_i} = D_{emp}(k, C_{min}; L, n)$, $D_{k_j} = D_{emp}(k, C_{max}; L, n)$. Assuming that, since h_{k_i} and h_{k_j} are relatively close (though not equal) we approximate $h_{k_i} \approx h_{k_j}$ (5). This simplification allows to compute g_{init} and h_{init} using only the empirical values and the estimated $V(L, n)$.

Then, we find the $V(L, n)$ value yielding the lowest average relative error by considering all pairs constructed from the mentioned candidate ranges: for each ($g_k(C; L, n)$, $h_k(C; L, n)$), the error in $D(k, C_i; L, n)$ is calculated for each C_i from the training set; this is performed for each value of k.

ii) Estimating Parameters $g_k(C; L, n)$ and $h_k(C; L, n)$. Following Sect. (3.1), for each value of k, although the values of $g_k(C; L, n)$ are relatively close for several *corpora*, they are not equal. Thus, $g_k(C; L, n)$ must be fine-tuned for each *corpus* size C in order to achieve accurate $D(k, C_i; L, n)$ predictions on a wide range of large *corpora* sizes. The same applies to $h_k(C; L, n)$. Given the estimated $V(L, n)$, the values of $g_k(C; L, n)$ and $h_k(C; L, n)$ for each training *corpus* C_{tr} are obtained in two phases: *Phase i)* Finding initial points ($g_{init}(C_{tr})$, $h_{init}(C_{tr})$) for this search: we apply (5) to obtain $g_{init}(C_{tr}) = g_{k_{i,j}}$ and $h_{init}(C_{tr}) = h_{k_{i,j}}$, with C_i and D_{k_i} instantiated to C_{tr} and $D_{emp}(k, C_{tr}; L, n)$ respectively, and C_j and D_{k_j} instantiated to the largest training *corpus* C_{max} and $D_{emp}(k, C_{max}; L, n)$. Assuming that, since h_{k_j} and h_{k_j} are relatively close (though not equal) we approximate $h_{k_i} \approx h_{k_j}$, thus quickly finding $g_{init}(C_{tr})$ and $h_{init}(C_{tr})$; *Phase ii)* Searching around $g_{init}(C_{tr})$ and $h_{init}(C_{tr})$ to find ($g_k(C_{tr}; L, n)$, $h_k(C_{tr}; L, n)$) minimizing the $D(k, C_{tr}; L, n)$ relative error.

iii) Estimating the Dependency Behavior of $g_k(C; L, n)$ and $h_k(C; L, n)$ for General Test corpora. To model the dependencies of $g_k(C; L, n)$ and $h_k(C; L, n)$ on C, for each k, allowing to calculate $D(k, C; L, n)$ (3) for each test *corpus*, we use *splines*, based on the training learned values. The *spline* implementation uses two hyperparameters, degree (dg) and smoothing level (s), denoted (dg_g, s_g) and (dg_h, s_h), respectively, for $g_k(C; L, n)$ and $h_k(C; L, n)$, tuned by cross-validation. Thus, to choose the quadruple that leads to the most accurate $D(k, C; L, n)$ predictions, a set of quadruple values (dg_g, s_g, dg_h, s_h) is generated such that $dg_g, dg_h \in \{3, 4, 5\}$ and $s_g, s_h \in \{0, 0.5, 1, 1.5 \ldots 7\}$. For each quadruple, leave-one-out cross-validation is employed: for each of N iterations (where N is the size of the training set, now used as cross-validation set), one *corpus* is used for validation, and the remaining $N-1$ *corpora* are used for training. Each iteration uses a different validation *corpus*.

In further detail, the training part of the cross-validation uses the specific values of $g_k(C; L, n)$ and $h_k(C; L, n)$ obtained for each *corpus*, and builds the *splines* for modeling $g_k(C; L, n)$ and for $h_k(C; L, n)$ *vs* C. These *splines* are then used to interpolate values of $g_k(C_{val}; L, n)$ and $h_k(C_{val}; L, n)$, where C_{val} is the out-of-one *corpus* for validation in each iteration of the cross-validation. Firstly,

$g_k(C_{val}; L, n)$ and $h_k(C_{val}; L, n)$ are used to obtain $D(k, C_{val}; L, n)$ prediction (using (3)) and the corresponding relative error. Secondly, we calculate the root mean square of the relative errors (in absolute values), considering all iterations of the cross-validation for the quadruple. Finally, the quadruple that shows the lowest error is chosen to model the dependencies of $g_k(C; L, n)$ and $h_k(C; L, n)$ on C using the full cross-validation set as training *corpora*. Then, $g_k(C; L, n)$ and $h_k(C; L, n)$ can be obtained from the *splines* for test *corpora*.

4 Maxima k values for Reliable $W(k, C; L, n)$ Predictions

To assess the predictions accuracy of $W(k, C; L, n)$ using the *corpora* test set, the range of k must allow reliable predictions. Due to the stochastic variability of empirical $D(k, C; L, n)$ and $D(k + 1, C; L, n)$, they can deviate from their means. Since $W(k, C; L, n) = D(k, C; L, n) - D(k + 1, C; L, n)$, such variations may significantly affect the empirical $W(k, C; L, n)$ values, specially when the means of $D(k, C; L, n)$ and $D(k + 1, C; L, n)$ are too close. For the empirical $W(k, C; L, n)$ to be reliable, the means of $D(k, C; L, n)$ and $D(k+1, C; L, n)$ must be sufficiently distant to reduce the probability of the corresponding empirical values being close. While the exact means are unknown, we use the empirical values from *corpora* as approximations. This introduces some risk, but obtaining large *corpora* to estimate these means is impractical. The results support this assumption (Sect. 5). Thus, let D_k and D_{k+1} follow Poisson distributions with $\lambda_k = D(k, C; L, n)$ and $\lambda_{k+1} = D(k + 1, C; L, n)$ as their means, reflecting the values for *corpora* with equal C, given L and n. Then, $W_k = D_k - D_{k+1}$, with variances $Var(D_k) = k$ and $Var(D_{k+1}) = k+1$. Since λ_k and λ_{k+1} are large enough, Normal distribution can be used. Hence, $W_k \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu = \lambda_k - \lambda_{k+1}$ and $\sigma^2 = \lambda_k + \lambda_{k+1}$. Thus, for W_k to be reliable we state that the relative deviation of W_k from μ should be smaller than a threshold ϵ (a small positive number) with a confidence probability P. Using the Cumulative Distribution Function ($\Phi()$) of the Normal distribution, we express the probability as: $P = 2\Phi\left(\frac{\epsilon\mu}{\sigma}\right) - 1$ leading to $\epsilon = \frac{\sigma}{\mu}\Phi^{-1}\left(\frac{P+1}{2}\right)$. Thus, for reliability, the relative deviation ϵ should satisfy:

$$\epsilon > \frac{\sqrt{\lambda k + \lambda_{k+1}}}{\lambda k - \lambda_{k+1}}\Phi^{-1}\left(\frac{P+1}{2}\right) \quad . \tag{6}$$

As k increases, D_k and D_{k+1} tend to become closer, leading to a maximum k value: the k-threshold for W_k reliability. As the difference between D_k and D_{k+1} increases with *corpus* size, larger *corpora* tend to have larger k-thresholds.

In summary, for a given $P = 0.95$ and $\sigma = 0.03$, the k-threshold of a *corpus* of size C, is the maximum k such that $0.03 > \frac{\sqrt{\lambda k + \lambda_{k+1}}}{\lambda k - \lambda_{k+1}}\Phi^{-1}\left(\frac{0.95+1}{2}\right)$ where λ_k and λ_{k+1} are the empirical $D(k, C; L, n)$ and $D(k + 1, C; L, n)$ values, respectively. This represents the maximum k for which the relative error of the prediction $W(k, C; L, n)$ using Eq. (4) is reliable, as it requires the empirical $W(k, C; L, n)$ value to be measured. The k-threshold values are in Sect. 5.

5 Results

5.1 The Corpora sets

We extracted files from *oscar-project* (https://oscar-project.org/) in order to build separate sample *corpora* collections: one in English and one in German. Files were randomly selected obtaining a collection of English *corpora* with the following sizes: 299 Mw, 365 Mw, 366 Mw, 5.48 Gw, 7.31 Gw, 11.3 Gw, 15.0 Gw, 20.5 Gw, 31.5 Gw, 40.6 Gw, 66.3 Gw, 82.7 Gw, 84.5 Gw, 101 Gw, 172 Gw and 373 Gw. For German, the following *corpora* collection was formed: 170 Mw, 281 Mw, 307 Mw, 1.23 Gw, 2.46 Gw, 4, 93 Gw, 9.85 Gw, 19.4 Gw, 24.3 Gw, 27.1 Gw, 39.1 Gw, 48.9 Gw and 52.2 Gw. To ensure that the test *corpora* sets covered sizes ranging from the magnitude of the smallest *corpus* to the magnitude of the largest one for each language, without sacrificing the size of the training sets, we selected *corpora* of the following sizes from the English collection: 366 Mw, 11.3 Gw, 31, 5 Gw and 82.7 Gw, 172 Gw and 373 Gw, and from the German collection: 307 Mw, 4.93 Gw, 24.3 Gw and 48.9 Gw. The remaining *corpora* were assigned to training or to cross-validation sets, depending on the needs.

To ensure fair counts while preserving text semantics, a space was added next to each of the following characters: {':', ';', ',', '(', ')', '[', ']', '<', ' >', '"', '!', '?'}. Inflected forms are counted as distinct words in *corpora*, affecting the estimated vocabulary sizes for each n-gram size. For all the *corpora* in those collections, n-gram counts ($1 \leq n \leq 6$) were performed for each *corpus*, except for the 172 Gw and 373 Gw English *corpora*, for which, due to the long computation times required, only the 1-gram counts are reported in this paper.

5.2 Experimental Results

The estimated vocabulary sizes for each n-gram size are: 2.95×10^9, 1.995×10^{10}, 3.335×10^{10}, 1.51×10^{11}, 2.48×10^{11} and $7,2 \times 10^{11}$, for English 1-grams, 2-grams,...,6-grams, respectively, and 1.80×10^9, 6.76×10^9, 2.48×10^{10}, 7.30×10^{10}, 2.20×10^{11} and 7.25×10^{11}, for German 1-grams, 2-grams,..., 6-grams, respectively. Note that these estimated vocabulary values are affected by the inclusion of all word inflections in the n-gram counting. The following values give an insight of the magnitude of the numbers of distinct n-grams: 3 579 008, 1 461 558 and 266 894 for the 1-gram of the 366 Mw English *corpus*, for $k = 1$, $k = 2$ and $k = 15$, respectively, and 7 560 504 911, 2 624 427 472, 317 955 329 for the 3-gram of the 82.7 Gw English *corpus*, for $k = 1$, $k = 2$ and $k = 15$, respectively.

Metrics for Evaluation For a language L and an n-gram size, let $D_{pred}(k, C)$ to be a prediction value obtained from $D(k, C; L, n)$, and $D_{emp}(k, C)$ to be the corresponding empirical value. Let $RED(k, C) = \left| \frac{D_{pred}(k,C) - D_{emp}(k,C)}{D_{emp}(k,C)} \right|$ represent the module of the Relative Error of that prediction. Similarly, $REW(k, C) = \left| \frac{W_{pred}(k,C) - W_{emp}(k,C)}{W_{emp}(k,C)} \right|$ gives the module of the Relative Error of a $W(k, C; L, n)$ prediction. The mean of $RED(k, C)$ for a given frequency k across a set of C

values is denoted as $MRED(k)$. The mean of $RED(k,C)$ for a given *corpus* size C over a set of k values is denoted as $MRED(C)$. Similarly, $MREW(k)$ and $MREW(C)$ are used for $REW(k,C)$. Also, $SRED(k)= \sqrt{\frac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}} RED(k,C)^2}$ represents the Root Mean Square of the Relative Error of the $D(k,C)$ predictions given k, across a set of C values. This measures the stability of $RED(k,C)$ wrt a given k along the set $\mathcal{C}$; the closer $SRED(k)$ is to $MRED(k)$, the more stable $RED(k,C)$ is for that specific k value. Also, $SRED(C)= \sqrt{\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} RED(k,C)^2}$ measures the stability of $RED(k,C)$ wrt C across a set of k values, where $\mathcal{K}$ is the set of k values used. Likewise, to measure the stability of $REW(k,C)$ for a specific value of k or C, we define $SREW(k)= \sqrt{\frac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}} REW(k,C)^2}$ and $SREW(C)= \sqrt{\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} REW(k,C)^2}$, respectively. Metric abbreviations are defined in the table captions.

Evaluating the Approaches Besides the proposed functions $D(k,C;L,n)$ and $W(k,C;L,n)$, we refer to the baseline model [11] as $D_b(k,C;L,n)$ and $W_b(k,C;L,n)$ (b denotes baseline). Also, to assess the isolated effect of the constancy assumption of $g_k(C;L,n)$ and $h_k(C;L,n)$ wrt the *corpus* size, we considered another approach denoted as $D_c(k,C;L,n)$ and $W_c(k,C;L,n)$, for evaluating the baseline model using cross-validation, instead of using the same *corpora* for training and testing as in [11]. Table 1 shows the mean relative errors, $MRED(k)$, for $k \in \mathcal{K}^{\mathcal{D}}= \{1,2,3\ldots,16\} \cup \{2^5,2^6,2^7,2^8,2^9,2^{10},2^{11},2^{12}\}$. The M^D column shows that the relative errors for $D(k,C;L,n)$ predictions are low across the entire range of k, with global means of 0.7%, 0.6% and 3.3% for 1-grams, 3-grams and 6-grams, respectively. Generally, for each k, M^D and S^D are relatively close, as the relative error remains stable across the different *corpora*. However, two outliers appear for 6-grams: 13.6% and 11.6% for $k = 2^{10}$ and $k = 2^{12}$, respectively. This is due to the relatively low empirical values for higher n-gram sizes and larger k values, becoming more sensitive to small variations.

By comparison, the significantly higher values in M_b^D show that the baseline is not able to handle such a large range of *corpora* sizes, as global means surpass 30% for these n-gram sizes. The baseline [11] does not present $D_b(k,C;L,n)$ values for $k > 16$. The M_c^D column shows that, after modifying the baseline to use cross-validation, while maintaining the assumption of constancy of $g_k(C;L,n)$ and $h_k(C;L,n)$ wrt C, the global means of the relative errors are still higher (reaching 16.8% for 6-grams) than those obtained by our approach (M^D). This highlights that the high relative errors of the baseline in a wide range of *corpora* are due to two issues: the inadequate estimation based on the same *corpora* set for both training and testing, and the constancy assumption of $g_k(C;L,n)$ and $h_k(C;L,n)$ wrt *corpora* sizes. Although not shown, 2-grams, 4-grams and 5-grams exhibit similar values.

Table 2 shows mean relative errors, $MRED(C)$ for the English *corpora*. For each case, all the $k \in \mathcal{K}^{\mathcal{D}}$ were used. The values of M^D and S^D are relatively close, showing that the $D(k,C;L,n)$ predictions have low relative errors and

Table 1: Mean relative errors, $MRED(k)$, for each k, denoted M^D, M_b^D and M_c^D, respectively, for models $D(k, C; L, n)$, $D_b(k, C; L, n)$ [11] and $D_c(k, C; L, n)$. $SRED(k)$ is represented by S^D. A global mean (GM) for each column (except S^D) is shown. Values shown for English test set (in percentage).

English

	1-grams				**3-grams**				**6-grams**			
k	M^D	S^D	M_b^D	M_c^D	M^D	S^D	M_b^D	M_c^D	M^D	S^D	M_b^D	M_c^D
1	**1.0**	1.3	87.4	4.5	**0.8**	1.0	20.6	6.0	**1.4**	1.6	27.6	1.7
2	**1.0**	1.1	72.0	3.3	**1.1**	1.3	35.0	0.4	**3.5**	4.2	46.0	2.7
3	**0.6**	0.7	67.7	3.2	**0.3**	0.4	35.3	1.0	**2.0**	2.5	40.1	5.2
4	**0.6**	0.9	64.1	3.0	**0.2**	0.3	35.3	1.0	**2.0**	2.5	40.1	5.2
5	**0.6**	0.8	61.3	2.9	**0.1**	0.1	35.7	1.3	**1.9**	2.3	38.2	5.6
6	**0.6**	0.8	57.9	2.8	**0.0**	0.1	35.7	1.8	**1.9**	2.2	37.1	4.3
7	**0.6**	0.7	56.6	2.7	**0.1**	0.6	36.6	2.6	**1.6**	1.8	34.5	2.7
8	**0.4**	0.5	54.1	2.6	**0.4**	0.6	35.6	3.2	**1.3**	1.5	33.5	5.1
9	**0.6**	0.7	53.0	2.4	**0.3**	0.4	35.7	4.2	**0.9**	1.2	30.5	10.0
10	**0.5**	0.6	52.1	2.3	**0.3**	0.4	35.3	5.0	**1.1**	1.2	29.2	13.5
11	**0.4**	0.6	51.4	2.2	**0.2**	0.3	35.2	5.9	**1.7**	1.8	31.9	18.0
12	**0.4**	0.6	50.1	2.2	**0.2**	0.3	35.2	6.5	**1.8**	2.0	33.5	20.7
13	**0.4**	0.6	49.5	2.2	**0.4**	0.5	35.2	7.0	**1.4**	1.9	35.0	22.6
14	**0.4**	0.6	48.3	2.1	**0.4**	0.6	35.3	7.5	**1.6**	2.1	36.1	25.1
15	**0.5**	0.6	47.5	2.1	**0.4**	0.5	35.3	8.1	**2.0**	2.6	38.4	28.7
16	**0.5**	0.6	46.6	2.1	**0.5**	0.5	35.5	8.5	**2.1**	2.8	39.6	30.1
2^5	**0.7**	0.8		2.1	**0.5**	0.6		12.4	**3.9**	5.3		37.1
2^6	**1.1**	1.7		2.5	**1.8**	2.7		12.2	**4.5**	5.4		21.8
2^7	**0.7**	0.8		3.7	**1.2**	1.5		12.9	**6.8**	9.5		17.8
2^8	**0.7**	0.9		4.5	**1.2**	1.2		14.2	**3.7**	6.2		32.8
2^9	**0.6**	0.8		4.8	**0.6**	0.9		16.7	**2.9**	3.5		46.7
2^{10}	**0.8**	1.2		4.1	**0.9**	1.4		17.9	**13.6**	15.5		26.1
2^{11}	**1.1**	2.0		3.6	**0.9**	1.2		19.3	**5.1**	6.3		9.2
2^{12}	**1.2**	2.0		4.2	**2.5**	4.6		19.4	**11.6**	17.2		10.7
GM	**0.7**		57.5	3.2	**0.6**		34.5	8.1	**3.3**		35.8	16.8

are stable for each *corpus* across the values of k. This is true for all n-gram sizes ($1 \leq n \leq 6$). The GM values range from 0.4% (2-grams) to 3.3% (6-grams). In comparison, the errors in M_b^D show significant errors for the baseline across the various test *corpora* and n-gram sizes, with *GM* values ranging from 18.0% (2-grams) to 57.5% (1-grams). The values of M_c^D show that the constancy assumption of $g_k(C; L, n)$ and $h_k(C; L, n)$ wrt C exhibits significant relative errors, namely for the smaller *corpora* in this large range set, reaching 54.7% (6-grams). For *corpora* sizes 172 Gw and 373 Gw, only 1-gram results are shown (Sect. 5.1).

Table 3 shows mean relative errors, *MRED*(C), for each of the German *corpora* set. $D_b(k, C; L, n)$ results for German are not included, since German is not reported in the baseline [11]. Again, $D(k, C; L, n)$ predictions generally show low values for the relative errors (M^D, S^D, and *GM*), similar to those obtained for English. Although for the smallest *corpus* (308 Mw), the error value is 8.2% (5-gram) and 9.7% (6-grams). Likely, these outliers could disappear if the training set were larger. In comparison to $D(k, C; L, n)$, for this *corpus* (308 Mw), the M_c^D approach reaches errors of 47.9% (5-grams) and 61.9% (6-grams).

Table 2: Mean relative errors, *MRED*(C), for each *corpus* C, denoted M^D, M_b^D and M_c^D, respectively, for models $D(k, C; L, n)$, $D_b(k, C; L, n)$ [11] and $D_c(k, C; L, n)$. *SRED*(C) represented by S^D. A global mean (*GM*) for each column (except S^D) is shown. Values shown for English test set (in percentage).

English

	1-grams				**2-grams**				**3-grams**			
C	$\boldsymbol{M^D}$	$\boldsymbol{S^D}$	$\boldsymbol{M_b^D}$	$\boldsymbol{M_c^D}$	$\boldsymbol{M^D}$	$\boldsymbol{S^D}$	$\boldsymbol{M_b^D}$	$\boldsymbol{M_c^D}$	$\boldsymbol{M^D}$	$\boldsymbol{S^D}$	$\boldsymbol{M_b^D}$	$\boldsymbol{M_c^D}$
366 Mw	**1.2**	1.4	31.0	8.5	**0.9**	1.2	34.7	13.6	**1.2**	2.2	43.3	26.7
11.3 Gw	**0.5**	0.6	61.7	2.0	**0.5**	0.8	18.4	1.3	**0.8**	1.4	36.7	2.4
31.5 Gw	**0.3**	0.4	66.4	2.0	**0.1**	0.2	12.6	1.2	**0.3**	0.4	31.8	2.6
82.7 Gw	**0.3**	0.5	66.0	0.4	**0.4**	0.4	6.2	0.3	**0.3**	0.4	26.1	0.6
172 Gw	**0.7**	0.8	62.5	2.1								
172 Gw	**1.2**	1.7	57.1	4.4								
GM	**0.7**		57.5	3.2	**0.4**		18.0	4.1	**0.9**		33.5	8.1
	4-grams				**5-grams**				**6-grams**			
366 Mw	**2.5**	5.6	32.4	36.7	**3.9**	7.9	22.3	47.3	**4.6**	9.0	24.4	54.7
11.3 Mw	**1.0**	1.5	41.2	3.5	**4.9**	5.8	35.6	6.1	**3.3**	4.1	35.8	6.3
31.5 Mw	**0.8**	1.7	40.2	3.5	**1.4**	2.5	39.9	3.0	**1.7**	2.9	39.7	4.7
82.7 Mw	**0.4**	0.7	38.5	0.8	**2.7**	4.1	41.5	0.9	**3.8**	5.9	43.3	1.4
GM	**1.2**		38.1	11.1	**3.2**		34.8	14.3	**3.3**		35.8	16.8

From Sect. 4, the reliability of $W(k, C; L, n)$ evaluation imposes restrictions on the k value. So, for each k for which $g_k(C; L, n)$ and $h_k(C; L, n)$ are trained, all training *corpora* should be used. Since $W(k, C; L, n)$ predictions should not apply to $k > k$-threshold, the k-threshold value for evaluating $W(k, C; L, n)$ is determined by the *corpus* with the smallest k-threshold, typically the smallest

corpus. The k-threshold values, given by (6), found for each n-gram size in each training set, are: 9, 16, 15, 15, 15, 15 for English 1-grams,...,6-grams, respectively, and 9, 15, 23, 17, 17, 17 for German 1-grams,..., 6-grams, respectively.

Table 3: Mean relative errors, $MRED(C)$, for each *corpus* C, denoted M^D and M_c^D, respectively, for models $D(k, C; L, n)$ and $D_c(k, C; L, n)$. $SRED(C)$ represented by S^D. A global mean (GM) for each column (except S^D) is shown. Values shown for the German test set (in percentage).

German

	1-grams			**2-grams**			**3-grams**			**4-grams**			**5-grams**			**6-grams**		
C (Gw)	M^D	S^D	M_c^D	M^D	S^D	M_c^D	M^D	S^D	M_c^D	M^D	S^D	M_c^D	M^D	S^D	M_c^D	M^D	S^D	M_c^D
.308	**2.3**	2.6	6.8	**2.3**	2.6	15.3	**3.2**	3.6	22.8	**5.8**	6.5	30.3	**8.2**	9.6	47.9	**9.7**	11.4	61.9
4.93	**0.3**	0.5	1.4	**0.8**	1.3	2.1	**1.4**	2.6	2.4	**2.0**	3.3	1.8	**3.1**	4.4	4.4	**3.6**	4.9	3.9
24.3	**0.2**	0.2	0.0	**0.2**	0.2	0.1	**0.2**	0.4	0.2	**0.3**	0.6	0.3	**0.6**	1.0	0.5	**0.5**	0.8	1.8
48.9	**0.1**	0.1	0.0	**0.1**	0.4	0.0	**0.2**	0.5	0.0	**0.3**	0.7	0.0	**0.6**	1.4	0.2	**0.5**	0.7	0.3
GM	**0.7**		2.1	**0.9**		4.4	**1.3**		6.4	**2.1**		8.1	**3.1**		13.2	**3.6**		17.0

Table 4 shows the mean relative errors, $MREW(k)$, for $k \leq k$-threshold for each n-gram size. M^W and S^W values, and their relative proximity, show, for all n-gram sizes, $W(k, C; L, n)$ predictions with low relative errors, stable for each k across the *corpora* set, with GM from 0.9% to 3.8%. In contrast, $W_b(k, C; L, n)$ predictions (M_b^W) show much higher relative errors: GM from 15.8% to 72.5% (1-grams). The 1-grams evaluation, by including the largest *corpora*, 172 Gw and 373 Gw, stresses the ability to handle large *corpora* scales. The high errors for 1-grams, by standing out from the other errors, highlight the baseline limitations.

For each English test *corpus* Table 5 shows the mean relative errors $MREW(C)$ for $k \leq k$-threshold and each n-gram size. M^W and S^W values, being relatively close, indicate stable $W(k, C; L, n)$ predictions across the k values, for all n-gram sizes, with GM from 0.9% to 3.8%. In contrast, M_b^W shows $W_b(k, C; L, n)$ predictions with much higher relative errors, GM reaching 72.5% for 1-grams. For German *corpora* set, Table 6 shows similar values for the $W(k, C; L, n)$ predictions relative errors. However, the outliers in M^W for 5-grams and 6-grams (9.1% and 12.2%) suggest that a larger training set could likely eliminate them.

Figure 2a compares $D(k, C; L, n)$ prediction curves with corresponding empirical values for the English *corpora* test set, for 2-grams and 3-grams. The curves overlap illustrates the low relative errors. Figure 2b shows $W(k, C; L, n)$ predictions, for 1-grams, and the corresponding empirical values, for the same test set, from 366 Mw to 373 Gw, for $k \leq k$-threshold. The curve overlap for each *corpus* reveals the low relative errors of the $W(k, C; L, n)$ predictions.

Table 4: Mean relative errors, $MREW(k)$, for each k, denoted M^W and M_b^W, respectively, for models $W(k, C; L, n)$ and $W_b(k, C; L, n)$ [11]. $SREW(k)$ is represented by S^W. A global mean (GM) for each column (except S^W) is shown. Values shown for the English test set (in percentage).

English

	1-grams			**2-grams**			**3-grams**			**4-grams**			**5-grams**			**6-grams**		
k	M^W	S^W	M_b^W	M^W	S^W	M_b^W	M^W	S^W	M_b^W	M^W	S^W	M_b^W	M^D	S^W	M_b^W	M^W	S^W	M_b^W
1	**1.4**	1.9	97.8	**1.6**	2.1	29.5	**1.3**	1.9	13.8	**2.0**	2.4	22.9	**2.2**	2.7	27.5	**2.2**	2.5	23.9
2	**1.9**	2.4	79.8	**2.9**	4.1	9.4	**2.1**	2.4	34.5	**2.3**	3.2	44.7	**4.5**	6.3	49.1	**5.0**	6.2	49.2
3	**0.5**	0.6	80.8	**0.6**	0.9	9.8	**0.6**	0.8	34.2	**1.1**	1.4	44.0	**2.0**	2.7	46.3	**2.0**	2.2	44.3
4	**1.1**	1.5	76.8	**0.7**	0.8	12.2	**0.9**	1.2	35.8	**1.1**	1.5	45.2	**2.9**	3.5	47.1	**3.0**	3.6	44.7
5	**0.6**	0.8	82.8	**0.3**	0.4	12.3	**0.3**	0.4	34.6	**0.6**	0.9	44.0	**2.1**	2.7	45.3	**1.9**	2.5	41.8
6	**1.4**	1.7	67.2	**0.5**	0.8	16.8	**0.5**	0.7	37.9	**1.0**	1.5	47.2	**2.7**	3.4	48.9	**2.7**	3.7	46.0
7	**1.8**	2.3	79.7	**1.0**	1.4	13.5	**1.5**	2.6	34.6	**2.8**	5.3	42.6	**4.8**	8.3	42.0	**6.8**	10.6	38.0
8	**1.3**	1.9	65.8	**0.8**	1.3	18.3	**1.0**	1.5	37.7	**1.7**	2.6	46.2	**2.2**	3.4	47.0	**4.6**	5.4	44.5
9	**1.2**	1.6	63.2	**0.3**	0.4	15.9	**0.3**	0.5	35.3	**0.5**	0.6	41.3	**0.9**	1.2	38.7	**2.5**	3.0	34.9
10			63.2	**0.5**	0.7	17.3	**0.7**	1.2	35.8	**1.3**	2.1	43.1	**2.7**	3.3	41.2	**5.0**	5.4	39.5
11			70.2	**0.3**	0.3	15.8	**0.3**	0.4	33.5	**2.9**	3.4	38.6	**1.8**	2.0	33.2	**3.7**	3.9	30.3
12			60.5	**0.9**	1.2	17.9	**1.5**	2.0	35.3	**6.2**	7.7	40.8	**4.8**	7.3	36.6	**7.0**	9.9	35.7
13			70.5	**0.5**	0.6	14.8	**0.5**	0.5	33.3	**1.0**	1.1	36.5	**2.0**	2.2	28.7	**2.7**	3.2	27.0
14			63.1	**1.1**	1.3	16.8	**1.9**	2.4	33.6	**3.7**	5.0	38.3	**4.9**	8.1	32.6	**6.8**	10.9	33.5
15			66.9	**0.6**	0.6	16.8	**0.8**	0.9	32.7	**1.3**	1.4	35.8	**1.9**	1.9	26.1	**1.3**	1.8	25.4
16				**1.0**	1.2													
GM	**1.2**		72.5	**0.9**		15.8	**1.0**		33.5	**2.0**		40.7	**2.8**		39.4	**3.8**		37.2

Table 5: Mean relative errors, $MREW(C)$, for each *corpus* C, denoted M^W and M_b^W, respectively, for models $W(k, C; L, n)$ and $W_b(k, C; L, n)$ [11]. $SREW(C)$ represented by S^W. A global mean (GM) for each column (except S^W) is shown. Values shown for the English test set (in percentage).

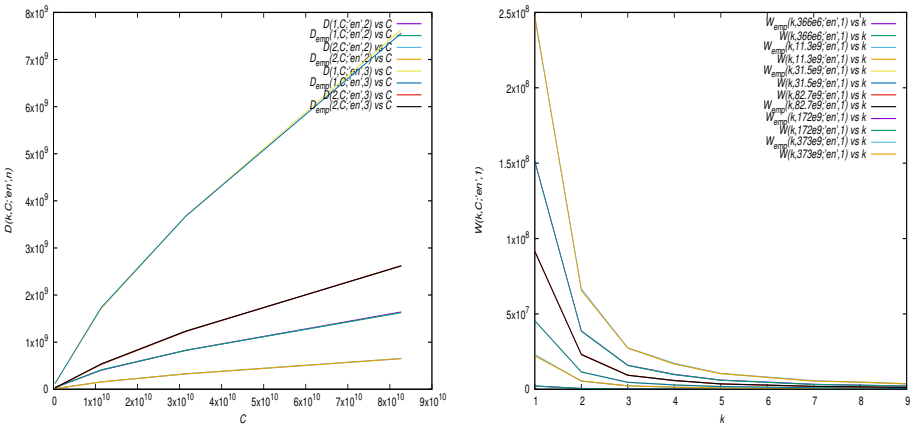
English

	1-grams			**2-grams**			**3-grams**			**4-grams**			**5-grams**			**6-grams**		
C (Gw)	M^W	S^W	M_b^W	M^W	S^W	M_b^W	M^W	S^W	M_b^W	M^W	S^W	M_b^W	M^W	S^W	M_b^W	M^W	S^W	M_b^W
.366	**2.3**	2.8	43.5	**1.7**	2.3	32.6	**1.9**	2.4	47.9	**4.1**	5.6	45.5	**5.6**	7.7	38.5	**7.2**	10.0	34.2
11.3	**1.1**	1.7	82.7	**1.0**	1.6	14.3	**1.2**	1.6	33.8	**2.1**	2.6	40.5	**2.8**	4.1	36.3	**3.6**	4.5	34.0
31.5	**0.4**	0.5	86.4	**0.3**	0.4	10.2	**0.4**	0.6	29.1	**0.6**	0.8	39.7	**0.7**	1.0	40.1	**0.9**	1.2	38.4
82.7	**0.6**	0.8	83.4	**0.4**	0.5	6.1	**0.3**	0.5	23.3	**1.1**	2.1	37.4	**2.2**	2.4	41.9	**3.6**	3.7	42.5
172	**1.2**	1.5	75.6															
373	**1.8**	2.1	63.5															
GM	**1.2**		72.5	**0.9**		15.8	**0.9**		33.5	**2.0**		40.7	**2.8**		39.4	**3.8**		37.2

Table 6: Mean relative errors, $MREW(C)$, for each *corpus* C, denoted M^W for model $W(k, C; L, n)$. $SREW(C)$ represented by S^W. A global mean (GM) for column M^W is shown. Values shown for the German test set (in percentage).

German

	1-grams		**2-grams**		**3-grams**		**4-grams**		**5-grams**		**6-grams**	
C (*Gw*)	$\boldsymbol{M^W}$	$\boldsymbol{S^W}$	$\boldsymbol{M^W}$	$\boldsymbol{S^W}$	$\boldsymbol{M^W}$	$\boldsymbol{S^W}$	$\boldsymbol{M^W}$	$\boldsymbol{S^W}$	$\boldsymbol{M^W}$	$\boldsymbol{S^W}$	$\boldsymbol{M^W}$	$\boldsymbol{S^W}$
.308	**2.4**	3.0	**3.6**	4.4	**4.5**	5.9	**6.3**	8.9	**9.1**	12.7	**12.2**	16.0
4.93	**0.9**	1.1	**0.8**	0.9	**1.8**	2.4	**1.1**	1.6	**2.4**	4.6	**7.3**	20.7
24.3	**0.4**	0.4	**0.2**	0.2	**0.2**	0.4	**0.2**	0.2	**0.3**	0.4	**0.5**	0.7
48.9	**0.1**	0.2	**0.1**	0.1	**0.5**	1.0	**0.1**	0.2	**0.3**	0.6	**1.0**	2.1
GM	**1.0**		**1.2**		**1.8**		**1.9**		**3.0**		**5.2**	



(a) (b)

Fig. 2: (a) Predicted and empirical values, $D(k, C; L, n)$ and $D_{emp}(k, C; L, n)$ versus C, for 2-grams and 3-grams, and $k \in \{1, 2\}$, in the English test *corpora* set. (b) Predicted and empirical values, $W(k, C; L, n)$ and $W_{emp}(k, C; L, n)$ *versus* k, for the 1-grams in the English test *corpora* set.

6 Conclusions

We aim to address the scalability issues raised by the need to predict the effect of the *corpus* size on n-gram frequency distributions when wide range of large natural language *corpora* are considered, encompassing sizes from hundreds of million words to hundreds of billion words. We propose a novel approach to handle the dependence of the model parameters on the *corpora* sizes. This was supported by a sound methodology for estimating the proposed model parameters, based on a state-of-the art for *corpora* training and testing, with cross-validation and generalisation through *spline*-based regression. Our goal is to achieve very low relative errors in the model predictions, and keeping them stable across the entire *corpora* size range. We focus on a prediction model applying uniformly to multi-words of different sizes, from 1-grams to 6-grams, considering the distribution of n-grams with low occurrence frequencies. In contrast to an approach assuming the parameter constancy wrt the *corpora* sizes, the conducted experimentation showed that the proposed approach led to very low relative errors (circa 2%) for the predictions of n-grams frequency distributions ($1 \leq n \leq 6$) in the range of low occurrence frequencies starting from 1 (singletons), and kept stable across a significant wide range of *corpora* sizes (from several hundred millions up to a maximum of 373 billion words in English), in two languages. This suggests that the proposed approach is promising to address the challenges posed by very large scale *corpora* sizes, and opens possibilities for handling relevant low occurrence multi-words in emerging and compelling applications, namely based on LLM.

Acknowledgments. This work has the financial support of FCT.IP: by UID/04516/ NOVA Laboratory for Computer Science and Informatics (NOVA LINCS); and by the FCT.IP project "Modelling the Statistical Distribution of n-grams in Large Natural Language Corpora", DOI 10.54499/2024.07024.CPCA.A1 .

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bernhardsson, S., da Rocha, L.E.C., Minnhagen, P.: The meta book and size-dependent properties of written language. CoRR, arxiv.org/abs/0909.4385 (2009)
2. 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)
3. Buck, C., Heafield, K., van Ooyen, B.: N-gram counts and language models from the Common Crawl. In: LREC'14. pp. 3579–3584 (2014)
4. Chierichetti, F., Kumar, R., Pang, B.: On the power laws of language: Word frequency distributions. In: ACM SIGIR Conference. pp. 385 – 394 (2017)
5. Goncalves, C., Silva, J., Cunha, J.: n-gram cache performance in statistical extraction of relevant terms in large corpora. In: Computational Science ICCS 2019 19th Intl. Conference. pp. 75–88. Lecture Notes in Computer Science, Springer (2019)
6. Ha, L.Q., Hanna, P., Ming, J., Smith, F.: Extending zipf's law to n-grams for large corpora. Artificial Intelligence Review **32**, 101–113 (2009)

7. Jones, E., T., O., et al., P.P.: SciPy: Open source scientific tools for Python. https://www.scipy.org/ (2001). https://doi.org/10.5281/zenodo.1913564
8. Lü, L., Zhang, Z., Zhou, T.: Deviation of zipf's and heaps' laws in human languages with limited dictionary sizes. Scientific Reports **3**(1082) (2013)
9. Newman, M.E.J.: Power laws, pareto distributions and zipf's law. Contemporary Physics **46**(5), 323–351 (2005)
10. Silva, J., Cunha, J.: An empirical model for n-gram frequency distribution in large corpora. In: Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020. pp. 840–851. LNCS, Springer (2020)
11. Silva, J., Cunha, J.: How large corpora sizes influence the distribution of low frequency text n-grams. In: Advances in Knowledge Discovery and Data Mining: PAKDD 2024. pp. 210–222. LNCS, Springer (2024)
12. Simon, H.: On a class of skew distribution functions. Biometrika **42**(3/4), 425 – 440 (1955)
13. Zipf, G.K.: Human Behavior and the Principle of Least-Effort. Addison-Wesley, Cambridge, MA (1949)