A statistical method for minimum corpus size determination

Assunta Caruso, Antonietta Folino, Francesca Parisi, Roberto Trunfio

Laboratorio di Documentazione - Department of Languages and Education,
Università della Calabria (Italy)
{assunta.caruso, antonietta.folino, francesca.parisi, roberto.trunfio}@unical.it

Abstract

A corpus is a well sized collection of structured text, (e.g. *articles*, *novels*, *legal documents*, *blog posts*, *oral transcriptions*, etc.) which is compiled based on specific goals. There is a growing interest in the use of corpora and therefore their constitution from both a qualitative and quantitative point of view is crucial. The attention in this paper, however, is put on *corpus size*. In particular, this paper presents an introductory study related to the determination of the minimum corpus size measured in terms of number of texts. The study focuses on a statistical technique commonly used for the determination of the sample size from a population with unknown size and variance. Measures of lexical richness are embedded in the proposed method to assess the quality. The technique is tested in the compilation of a specialist corpus for tourism in Italy. Findings provided by our numerical results suggest that the proposed statistical technique is worth further investigation so that it can be used as a standard decision support tool in corpus size definition.

Riassunto

Un corpus può essere definito come una raccolta di testi strutturati (per es. *articoli, romanzi, testi legali, post pubblicati su blog, trascrizioni di testi orali, ecc.*), costituita secondo specifici criteri. La costituzione di un corpus da un punto di vista quali-quantitativo assume un ruolo cruciale, in particolare alla luce del crescente interesse nell'uso dei corpora. Nel presente lavoro l'attenzione è posta sulla problematica di stimare la dimensione di un corpus. Nello specifico, questo lavoro presenta uno studio introduttivo concernente la determinazione della dimensione minima di un corpus misurata basandosi sul numero di testi che lo compongono. A tale scopo, l'attenzione è posta su una tecnica statistica utilizzata per la determinazione della dimensione di un campione a partire da una popolazione di dimensione e varianza non note. Al fine di accertare la bontà del metodo proposto alcune misure di ricchezza lessicale sono utilizzate. Tale tecnica è infine testata nella costruzione di un corpus specialistico per il comparto del turismo in Italia. I risultati numerici riportati nel presente lavoro suggeriscono che la tecnica qui proposta è meritoria di essere ulteriormente approfondita al fine di determinarne un suo uso quale strumento di supporto alle decisioni nella definizione della dimensione di un corpus.

Keywords: corpus size, quantitative approach, texts, statistics

1. Introduction

Corpus linguistics is the science related to the compilation and analysis of a collection of texts, i.e. *a corpus*, which can be defined as "*a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 2004).* Hence, a corpus is the basic resource for corpus linguistics and therefore its compilation is a key step in the study of languages (Hunston, 2006). (Cortelazzo and Tuzzi, 2008) suggest that the compilation of a corpus is carried out using personal experience and statistical methods.

Although during the past decades a great deal of research effort has been spent in studying the compilation of corpora (McEnery and Wilson, 2001), some questions are still open and are worth discussing. In particular, from the point of view of statistical methods, an interesting issue is related to the evaluation of corpus size. A common unit of measure for corpus size used by several authors is the total *number of tokens* (i.e. words) that composes the corpus (Hunston, 2006). Another useful unit of measure is the *number of texts*, which can be correlated to the former (e.g. by using the average number of tokens per text). In our perspective, the latter measure includes some practical aspects which the linguist must face (e.g.: How many articles must be bought ? How many speeches should be transcribed ?).

Thus, this paper aims at identifying a statistical method to be used for the determination of the minimum number of texts to be selected to compile a sample corpus. In the following, we assume that the total corpus has an unknown size or that its size is difficult to be estimated. This assumption is true especially when the corpus aims at collecting texts from an unmanageable variety of sources (e.g. *blog posts*, *newspaper articles*, etc.).

A review of the literature shows how some studies assert that the minimum corpus size can be estimated *a priori* by resorting to specialized algorithms (Lauer, 1995); other studies express skepticism toward generalized and automated methods designed for corpus size determination and thus suggest calculating the corpus size *a posteriori* or incrementally (De Haan, 1992 ; Yang et al., 2000). More specifically, a common practice is to apply the methods used to study the growth of the vocabulary of a text (Baayen, 1996) to the whole corpus. Hence, these studies are devoted to estimating a lower bound on the number of tokens required to define a corpus. For instance, (Yang et al., 2000, 2002) proposed a predictive method based on the piecewise curve-fitting algorithm. This method is based on the frequency of lemmas in the analyzed corpus. As remarked by Yang et al., the predictive capabilities of the method are weak, since it overestimates the number of tokens to be used in a corpus (e.g. *200 million tokens per corpus* vs *1 million tokens* suggested to be used by (Tognini Bonelli and Sinclair, 2006)). Another approach proposed by (Pastor and Domínguez, 2007) proposed to incrementally augment the number of texts in the corpus and then to iteratively evaluate the value of a desired measure of lexical richness. Their method stops once the curve related to the proposed measure is saturated. A more sophisticated method is proposed by Juckett (2012) who suggests using a probabilistic method to confirm *a posteriori* the minimum size of a corpus of clinical texts with respect to a known population of texts. In our opinion, the major weaknesses of the method proposed by Juckett are: *(i)* the need for a comparison corpora containing appropriate common word usage; *(ii)* the need for full texts to be collected.

Accordingly, no practical and rigorous method is yet available as *de facto* standard in corpus linguistics to define the size of a corpus in terms of number of texts. Hence, in Section 2 we investigate some statistical techniques used in sample size determination aiming to define a general approach to be used in corpus linguistics. Then, in Section 3 we propose a case study to assess the effectiveness of the proposed technique. Finally, in Section 4 we draw conclusions and we comment on further research.

2. A statistical method to estimate minimum corpus size

In statistics, a sample is a collection of data extracted from a population using a specific procedure. In any empirical study, the determination of the sample size covers a key role in the process of making inference from the overall population. Statisticians have to deal with the trade-off between the expense of collecting large amounts of data and ensuring the quality of the inference. The analogy with the activity of corpus compilation is straightforward: the population is the whole production of written and oral texts in the particular field of study and the sample is the corpus to be defined.

Therefore, in the following sections we report some necessary notations and equations about sample size determination in statistics and then we explain how these techniques can be applied in corpus size determination.

2.1. Sample size determination

Let *n* be the sample size, *1-α* the desired confidence level, $z\alpha/_2$ the normal quantile at the desired confidence level, σ^2 the true variance for the population and ε the maximum desired marginal error (such that ε>0). Roughly speaking, if the sample size *n* is calculated as follows (Woods et al., 1986):

$$n = \left\lceil \left(\frac{z\alpha/_2 \cdot \sigma}{\varepsilon} \right)^2 \right\rceil \tag{1}$$

then an error at ±ε with respect to the expected mean of a given index is guaranteed with a probability *1-α*. Eq. (1) is based on the assumption that the true variance σ^2 is known *a priori* for the target population. This assumption is not mandatory *per se*, since some approaches allow us to replace σ^2 with a *target variance* or with the variance of a reference sample if available.

However, when dealing with a specialized corpus, formulating a hypothesis on the value of the true variance for the corpus could be a hard task and sometimes counterproductive. Moreover, not even by replacing σ^2 with the unbiased estimate of the variance S^2_{n-1} can the problem be solved, since S^2_{n-1} must be calculated by resorting to all the *n* texts selected for the corpus.

In order to measure *n*, here we propose to use an effective and consolidated statistical parametric method. The method is based on the use of an initial sample of size n_0 (with $n_0 \leq n$) to establish if the initial sample size is big enough or, if otherwise, how many additional texts must be sampled. The method is derived from the *Rinott Procedure* (Rinott, 1978 ; Chen, 2011), a statistical method for indifference zone selection that embeds a slightly different version of eq.(1).

The procedure is as follows. First, identify a sample corpus composed by a collection of n_0 texts that are properly selected (all the necessary remarks about this point are in the following subsection). Then, for each text *t* in the sample corpus, with $t=1,\dots,n_0$, let I_t be an index of the lexical richness of text *t*. Observe that all the measured indices are independent under the assumption that the texts are suitably selected. Successively, measure the unbiased estimate of the sample variance $S^2_{n_0-1} = \frac{1}{n_0-1}\sum_{t=1}^{n_0}(I_t - \bar{I})^2$ where $\bar{I} = \sum_{t=1}^{n_0} I_t/2$ is the sample mean of the selected index. Thus, the minimum corpus size with the desired error and confidence level can be measured as follows:

$$n = max\left\{ n_0,\ \left\lceil \left(\frac{z\alpha/_2 \cdot S_{n_0-1}}{\varepsilon} \right)^2 \right\rceil \right\} \tag{2}$$

Clearly, the additional number of texts added to the sample corpus is *n* - n_0. For the sake of completeness, figure 1 illustrates a synthetic flowchart of the steps of the proposed procedure for determining the corpus size.

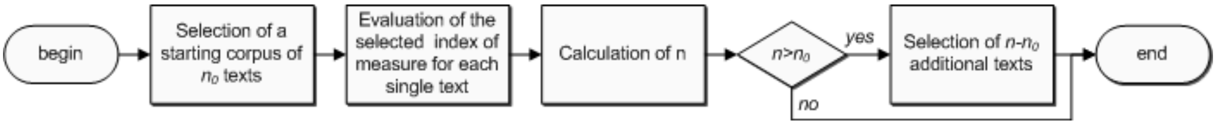
begin
Selection of a starting corpus of n_0 texts
Evaluation of the selected index of measure for each single text
Calculation of n
$n>n_0$
yes
no
Selection of $n-n_0$ additional texts
end

Figure 1. The figure shows a synthetic flow-chart of the proposed procedure for the determination of the minimum corpus size.

2.2. Evaluating the index of lexical richness

In order to implement the procedure described in the previous section, an index for measuring the lexical richness of each text in the sample corpus must be selected. In the literature, an extensively used measure is the *type-token ratio* (TTR) (Hardie and McEnery, 2006).

Given a text *t*, let N_t be the number of tokens in *t* and V_t be the number of types in *t*, then the simplest measure for the TTR of text *t* is:

$$TTR_t = \frac{V_t}{N_t} \tag{3}$$

Observe that the measure in eq.(3) is a number defined in [0,1], since for any text results $1 \leq V_t \leq N_t$. Some interesting attempts to improve the TTR index have been proposed in the literature (Guiraud, 1954 ; Herdan, 1960 ; Brunet, 1978 ; Ejiri et Smith, 1993 ; Covington et McFall, 2010), but only a few of these variants possess some key properties that candidate them to be used in our text comparison. In the following, a TTR variant proposed by Herdan (1960) and usually addressed as *Herdan's C* or *LogTTR* is reported here for a text *t*:

$$LogTTR_t = \frac{\ln V_t}{\ln N_t} \tag{4}$$

defined in [0,1] under the additional condition that if $V_t = 1$, then it must result that $N_t > 1$. The variant by Herdan is considered here since, like the original measure, it allows us: *(i)* to attain a non-indeterminate value of the measure for any V_t and N_t in the domain and *(ii)* to assign at a lower value of the index a lower degree of lexical richness (and *vice versa*). An example of the use of both indices on a text is reported in figure 2. In particular, the plots in figure 2 report the value of the index *(y* axis) measured after a specific number of tokens (*x* axis).

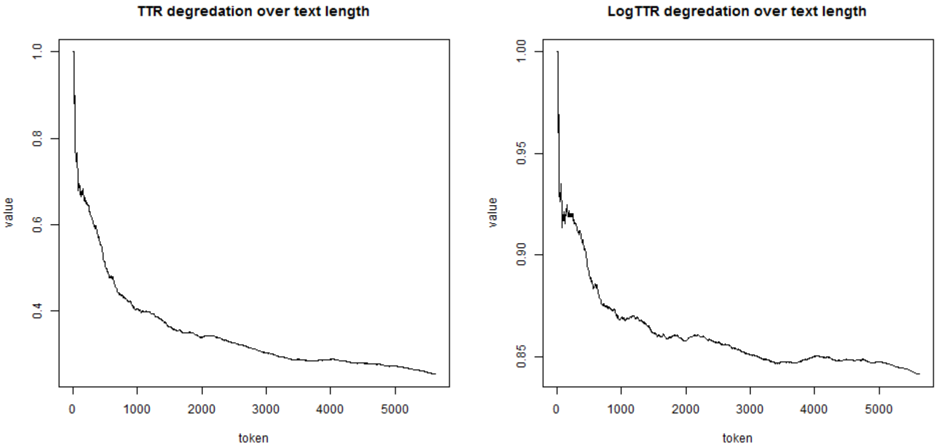

Figure 2. The plot for the classic TTR (on the left) and the LogTTR (on the right) for the Italian law "Legge 29 marzo 2001, n. 135 – Riforma della legislazione nazionale del turismo" are given for illustrative purposes.

For a better estimate of the corpus size, we suggest repeating the calculation of eq.(2) once for each of the indices in eqs.(3) and (4), assuming as the estimated value of n the larger one returned by (2).

2.3. Choice of the initial sample

As mentioned in Section 2.2, here we observe that the sample corpus must be defined by turning to a collection of carefully selected texts. Apart from some key aspects related to corpus quality, which have to be dealt by the linguist by focusing on the objective of the corpus (Biber et al., 1998), here we first provide some remarks on a methodological guideline related to the qualitative structure of the corpus and then we state some necessary quantitative conditions that must be met in order to apply the statistical method illustrated in this paper.

Generally speaking, a corpus is a collection of heterogeneous texts (e.g. laws, novels, press, etc.), which usually can be partitioned in subsets of "homogeneous" texts, which can be assumed to be particular text categories. For instance, the well-known Brown corpus is a collection of 500 sample texts arranged in 15 categories (Francis and Kucera, 1979). Therefore, to provide a better estimate of the overall corpus size *n*, we suggest identifying the categories that must compose the corpus and thus to turn to stratified sampling. Specifically, a category can be defined as the set of texts that shares a predetermined *level of similarity* of the terms identified as meaningful according to the purposes of the categorization (Guarasci, 2008).

Once C categories are identified by the domain expert, then the number of texts to be added to each category must be defined. Formally, let c be a category (*c=1,…,C*) and n^c be the number of texts in category c, then we can state that:

$$n = \sum_{c=1}^{C} n^c \tag{5}$$

The n^c can be obtained by using eq.(2) by selecting an initial number of texts n_0^c. Once a category has been suitably defined, the selection of n_0^c texts should be an easy task for the linguist. To use eq.(2) we suggest selecting at least 30 texts, while in practice the larger the number of initial texts, the better the variance estimate.

This approach allows us to apply the proposed method to texts with a level of similarity fixed *a priori* and thus mitigating the effect on the sample variance usually due to the calculation of the TTR on texts from different categories. This observation can be easily ascertained by considering that each text type has generally a specific distribution of lexical richness, which depends on language, time, etc. (Van Gijsel et al., 2006; Malvern et Richards, 2012).

To select the n_0^c texts for the initial sample of the c-th category, the following condition must hold: *(i)* each text must have a minimum number of tokens, called N_{min}, such that a sufficient lexical richness can be guaranteed; *(ii)* for each text t, the corresponding index of lexical richness (i.e. the TTR_t and $LogTTR_t$) is calculated at the first N tokens, such that $N_{min} \leq N \leq N_{max}$, where $N_{max} = min_{t=1,\ldots,n_0^c}\{N_t\}$ is the smallest number of tokens across all the texts in the initial sample of category c and N_t is the total number of tokens in text t. The latter condition is due to the fact that the TTR-value depends on the length of the analyzed text and therefore the comparison of different values makes sense at the same number of tokens.

To remove the aforementioned condition on the measure of the TTR indices, some of the statistical approaches developed for the *regenerative method* can be used (Crane and Lemoine, 1977; Hillier and Lieberman, 2010). These statistical techniques can be used to embed the length of the text in the desired measure; the main drawback is that these approaches introduce some additional effort in the computation of the sample variance due to the given correlation between the length of the text and the TTR index. The description of these techniques is out of the scope of this introductive study, hence we refer the interested reader to the literature.

3. Computational experiments

We conducted a series of computational experiments to assess the applicability of the proposed statistical methodology in corpus linguistics. To this extent, experiments have been carried out on selected categories of specialized texts. Specifically, the texts are part of a corpus currently being compiled for the construction of a tourism thesaurus at the Laboratorio di Documentazione (*LabDoc*), Department of Languages and Education, Università della Calabria, Italy. The activity is conducted in the context of the project DiCeT–INMOTO-OR.C.HE.S.T.R.A[1], part of the "Programma Operativo Nazionale Ricerca e Competitività 2007 -2013 – Smart Cities and Communities and Social Innovation", funded by the Italian *Ministero dell'Istruzione, dell'Università e della Ricerca* (MIUR). The computational experiments have been conducted by using *koRpus* (Michalke, 2013), an *R* package for text analysis that makes use of the *TreeTagger* tool (Schmid, 1994 ; Schmid, 1995).

3.1. Minimum corpus size determination for the Italian laws and regulations in the tourism domain

This group of experiments aims at identifying the number of laws/regulations emitted by both the central and regional Italian governments related to the domain of tourism. To demonstrate the effectiveness of the proposed statistical method, in these experiments a known and available population of selected texts is considered. Obviously, the whole population of texts can be analyzed in order to provide some upper bound values (e.g. the *overall number of types*) to be used for comparison with the statistics related to any sample corpus drawn from the total population. In particular, 120 texts (national laws, regional laws and other regulations) have been selected. According to TreeTagger, the texts have from about 2,000 to 50,000 tokens (6,800 tokens on average) and the overall number of tokens and types for the whole corpus are 824,903 and 25,645, respectively. Three experiments, called *A1*, *A2* and *A3*, are described here. Each experiment consists of 100 batches of texts (or sample corpora), where each batch is composed of 50, 75 and 100 texts randomly sampled from the set of 120, respectively for experiments *A1*, *A2* and *A3*. Therefore $n_0=50$, $n_0=75$ and $n_0=100$, are set for *A1, A2* and *A3,* respectively. Then, for each batch n has been calculated by referring to eq.(2) by setting $1\text{-}\alpha=0.995$ and ε=0.01. The former fixes the normal quantile at $z_{\alpha/2}$=2.58. The latter ensures a very small width of the confidence interval with respect to any estimate of the true mean TTR (or LogTTR) value and therefore a high accuracy of the estimate (specifically ensures that the true mean contains a total width of 0.02 with a probability 0.995). Finally, the N=N_{max} has been set.

For reasons of space, only some synthetic statistical data for the experiments *A1*, *A2* and *A3* are reported in Table 1. For N, the minimum, mean, maximum and standard deviation measured across all the 100 batches are reported. For each batch, the sample mean and sample standard deviation for the TTR and LogTTR are calculated during the experiments.

[1] DiCeT – LivingLab Di Cultura e Tecnologia; INMOTO – INformation and MObility for TOurism; OR.C.HE.S.T.R.A. – ORganization of Cultural HEritage for Smart Tourism and Real-time Accessibility.

Starting from these measures, Table 1illustrates: the grand mean (i.e. the mean of the 100 means); the minimum and maximum of the sample means; the minimum and maximum standard deviation of the batches. Finally, by using eq.(2), the minimum, mean, maximum and standard deviation measured across all the 100 batches are reported for the calculated final number of texts *n*.

	TTR					***LogTTR***				
Exp.	*min*	*g.mean*	*max*	*min std. dev*	*max std dev*	*min*	*g.mean*	*max*	*min std. dev*	*max std dev*
A1	0.312	0.321	0.328	0.030	0.046	0.847	0.851	0.854	0.012	0.020
A2	0.320	0.327	0.335	0.028	0.042	0.851	0.854	0.856	0.011	0.018
A3	0.323	0.328	0.332	0.029	0.036	0.852	0.854	0.855	0.012	0.014

	N				***n***			
Exp.	*min*	*mean*	*max*	*std.dev*	*min*	*mean*	*max*	*std.dev*
A1	2,060	2,126.0	2,240	28.31	60	102.2	144	198.08
A2	2,105	2,116.0	2,235	18.40	76	101.1	129	89.00
A3	2,105	2,109.3	2,135	9,56	100	100.0	100	0.00

Table 1. The table reports the numerical results for N, n and the TTR and LogTTR indices in experiments A1, A2 and A3.

According to Table 1, in all three experiments the TTR and LogTTR values are measured at about 2,100 tokens starting from the beginning of each text. In our experiments, the TTR index always provides more variance with respect to the LogTTR index; to highlight this observation , note that the worst case variance for the LogTTR (last column in Table 1) is half the best case variance for TTR (fifth column in Table 1), i.e. 0.020^2 vs 0.030^2 for *A1*, 0.018^2 vs 0.028^2 for *A2* and 0.014^2 vs 0.029^2 for *A3*. Thus, the largest value for *n* is always returned by using the variance of the TTR indices in eq.(2).

The more interesting result is related to the minimum number of texts returned by using eq.(2). In fact, the average value of *n* is 102, 101 and 100 for experiments *A1*, *A2* and *A3*, respectively. Hence, it seems that the variance estimator of the TTR value embedded in eq.(2) lets *n* converge to a common value, i.e. a minimum number of about 100 texts. Using the data from Table 1 the confidence intervals for *n* can be calculated. In particular, the 99% confidence interval for *n* in the experiments *A1*, *A2* and *A3* is, respectively, [98.57; 105.83], [97.7; 102.56] and [100; 100]. We can conclude that 100 seems to be a sufficient number of texts for the laws and regulations in the domain of tourism.

Henceforth, we assume that 100 is the minimum number of texts that can be used to compile a sample corpus for the population at hand. To test this hypothesis, some additional measures related to the three experiments are reported in Table 2. Initially, for each batch the total number of types *V* is calculated across all the texts collected at the first stage of the procedure illustrated in Section 2 (i.e. for the first n_0 texts of each batch). Using these measures, the gap between the total number of types in the corpus (i.e. 25,645) and the number of types in each batch, called gap_{types}, is obtained. Then, starting from these values, for each experiment the minimum, mean, maximum and standard deviation for *V* and the minimum, mean, maximum and standard deviation for gap_{types} is reported in Table 2.

Observe that, according to the aforementioned n_0 values, the batches in experiments from *A1* to *A3* include about 40%, 60% and 80% of the texts of the whole population, respectively for experiments *A1, A2* and *A3*. In the light of this observation, a better reading of the gap_{types}

measure can be provided. In particular: sampling 40% of the texts (experiments *A1*), the average gap is 0.38; sampling 60% of the texts (experiments *A2*), the gap reduces to 0.21; finally, sampling 80% of the texts as in *A3*, the gap is 0.08. In our opinion, the last gap (about 8% of types lost) could confirm that the minimum number of texts returned by eq.(2) provides a good sample corpus with respect to the number of types being captured.

	V				***gap*** $_{types}$			
Exp.	*min*	*mean*	*max*	*std.dev*	*min*	*mean*	*max*	*std.dev*
A1	13,918	15,990	19,097	1,246.5	0.26	0.38	0.46	0.049
A2	17,186	20,343	22,322	977.9	0.13	0.21	0.33	0.038
A3	22,036	23,496	24,644	610.1	0.04	0.08	0.14	0.024

Table 2. The table reports the numerical results for V and gap_{types} *in experiments A1, A2 and A3.*

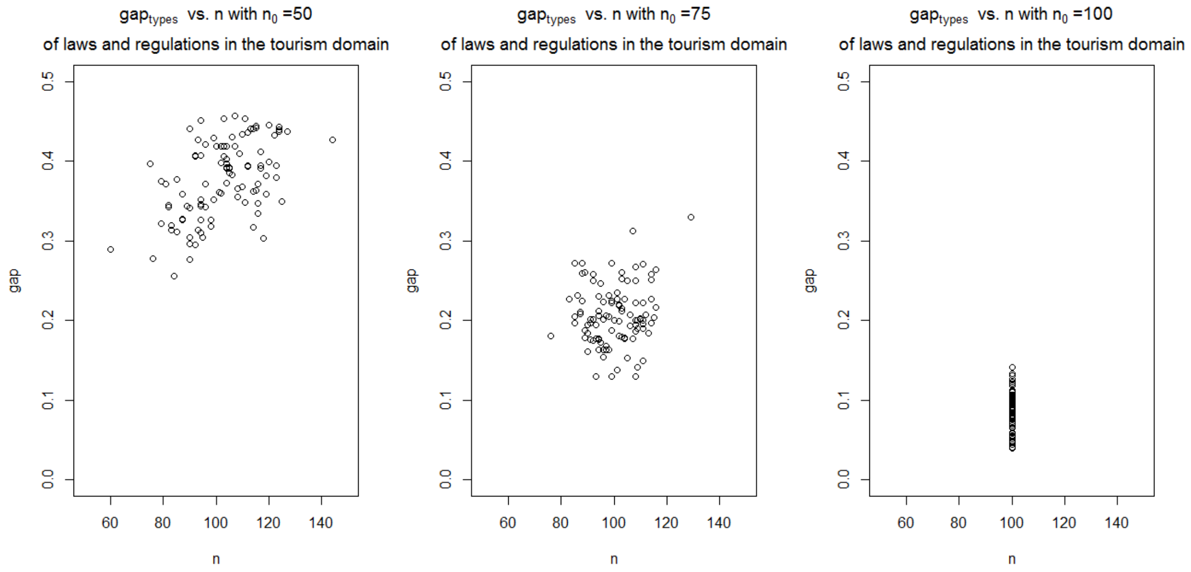

Figure 3. The three plots show the couple of values reported for each batch corresponding to the number of texts n returned by eq.(2) (x axis) and the gap between the number of types in a batch and the number of types in the whole population (y axis), respectively for n_0*=50, 75 and 100.*

In particular, figure 3 illustrates the plots related to the couple of values for *n* and gap_{types} observed for the batches in the experiments *A1*, *A2* and *A3* (shown in the first, second and third plot, respectively). In particular, each couple *n*-gap_{types} is related to one batch-out-of-100 and is depicted in a plot as a small empty circle. As one may observe, the smaller n_0 is , the more widely spread the circles are. In fact, for n_0=50, the suggested *n* ranges from 60 to 144 and gap_{types} from 0.26 to 0.46 (see Table 1 and 2, respectively). When n_0 increases to 75, *n* ranges from 76 to 129 and gap_{types} from 0.13 to 0.33. Hence, the points are more grouped for n_0=75 than for n_0=50, reflecting the fact that lower values of the initial sample provide a more biased estimation of the index of interest and thus less precise predictive power. For n_0=100 we have n=100 and the gap_{types} ranges from 0.04 to 0.14. Clearly, the value for n_0 should be the largest number with respect to a fixed cost-time budget. For a very large n_0 all the circles converge to a single point located on the *x*-axis (e.g., in our corpus this phenomenon occurs for n_0=120).

3.2. Minimum corpus size determination for the Italian journals in the tourism domain

We have identified a set of 220 journals published in Italy for the domain at hand in the period from January 1, 2008 to July 20, 2013. The estimated number of periodic publications for this set of journals is more than 4,000. Most of them are not available in electronic format, thus they must be digitized by resorting to the time consuming activity of image scanning; subsequently, a software for *optical character recognition* (OCR) must be used to convert the scanned images into texts. In addition, compiling a corpus of specialized journals by collecting the whole population would require a relevant amount of money related to the journals' purchase cost and workforce employed in the aforementioned manual activities. Both these issues allow us to insist on looking into the study of a statistical methodology capable of assessing if the number of selected texts is enough or, otherwise, how many texts must be sampled.

Therefore, here we present a second group of experiments that focuses on three categories of Italian journals from the domain of tourism. The goal is to identify the number of journals from this set of journal categories to be used in the corpus composition. In particular, we focused on the following three correlated categories: *professional training*; *business travels*; *destinations*.

In a similar fashion to what has been done in Section 3.1, three experiments, called *B1*, *B2* and *B3*, are reported here. They consist in selecting 100 batches composed of 75, 115 and 150 texts sampled from a set of 300 texts for *B1*, *B2* and *B3*, respectively. *Ergo* n_0 has been set at 75, 115 and 150. The set of 300 texts has been acquired from the whole population of periodic publications related to the Italian journals in the domain of tourism, according to criteria established by a group of content specialists. Since the estimated size of the whole population is very large, we expect a more reliable estimate of *n* and thus in eq.(2) we set $1\text{-}\alpha=0.9999$, that fixes $z_{\alpha/2}$=3.72, and again ε=0.01. According to TreeTagger, the 300 selected texts have from about 1,800 to 40,000 tokens (9,200 tokens on average) and the overall number of tokens and types for all the 300 texts are 3,175,281 and 133,442, respectively.

	TTR					***LogTTR***				
Exp.	*min*	*g.mean*	*max*	*min std. dev*	*max std dev*	*min*	*g.mean*	*max*	*min std. dev*	*max std dev*
B1	0.447	0.456	0.470	0.029	0.045	0.893	0.896	0.899	0.009	0.013
B2	0.452	0.458	0.464	0.033	0.043	0.894	0.896	0.898	0.010	0.013
B3	0.454	0.458	0.463	0.034	0.043	0.894	0.896	0.897	0.010	0.013

	N				***n***			
Exp.	*min*	*mean*	*max*	*std.dev*	*min*	*mean*	*max*	*std.dev*
B1	1,850	1,905.2	2,080	71.97	115	201.1	276	37.70
B2	1,850	1,885.6	2,030	44.66	155	201.7	258	28.87
B3	1,850	1,871.1	1,910	23.95	164	201.6	254	18.61

Table 3. The table reports the numerical results for N, n and the TTR and LogTTR indices in experiments B1, B2 and B3.

The same synthetic statistical data given in the previous section for *TTR*, *LogTTR*, *N* and *n* are calculated here for the experiments *B1*, *B2* and *B3* and are reported in Table 3. Comparing the results for the TTR and LogTTR indices, the observation reported in the previous section regarding the variance provided by the two indices is confirmed: the TTR always provides more variance with respect to the LogTTR index and so the largest value for *n* is always returned by using in eq.(2) the variance of the TTR indices. The TTR and LogTTR indices are evaluated, on average, at 1,900, 1,890 and 1,870 tokens from the beginning of the texts for batches in experiments *B1*, *B2* and *B3*, respectively. We recall that the number of tokens *N* is calculated in accordance to the N_{max} value (see Section 2). Thus, the larger n_0, the greater the probability to get a text *t* with a small N_t and so to get a smaller *N*.

We highlight that the same interesting result obtained in experiments *A1, A2* and *A3* is repeated here in experiments *B1, B2* and *B3*. In particular, the average value returned for *n* is 201.1, 201.7 and 202.6 for B1, B2 and B3, respectively. Then, it seems again that the TTR-based implementation of eq.(2) lets *n* converge to a minimum number of about 202 texts in all three experiments.

Hence, in the following we assume that *n*=202 is a sufficient number of texts to be sampled for the corpus which is the object of our study. Clearly, there is no upper bound on the number of types from the whole population to be used as reference point. Nonetheless, at least the value of gap_{types} for the batches with n_0=75, 115, 150 (i.e. for experiments *B1*, *B2* and *B3*) and the corresponding final samples with *n*=202 texts can be measured with respect to the available collection of 300 texts. These measures enable us to compare the gain in terms of additional types from n_0 to n and also to ascertain the gap by taking into account only 202-out-of-300 texts.

To this extent, the total number of types *V* per each batch of 75, 115, 150 and 202 texts is initially calculated. Finally, some statistical measures for *V* and gap_{types} are reported in Table 4.

	V				gap_{types}			
# texts	*min*	*Mean*	*max*	*std.dev*	*min*	*mean*	*max*	*std.dev*
75	29,692	31,642	33,259	1,016.6	0.75	0.76	0.78	0.008
115	50,193	52,730	54,833	1,322.6	0.59	0.60	0.62	0.010
150	86,837	88,857	90,866	1,204.1	0.32	0.33	0.35	0.009
202	119,872	122,098	123,944	1,160.7	0.07	0.09	0.10	0.009

Table 4. The table reports the numerical results for V and gap_{types} *for batches with 75, 115, 150 and 202 texts.*

As a matter of fact 75, 115, 150 and 202 texts form 25%, 38%, 50% and 67% of all the available texts, respectively. In light of this observation and referring to the numbers in Table 4, some comments are provided below.

From Table 4 one may notice that the average gap_{types} decreases with the number of texts in a non-linear fashion. In particular, for batches collecting 75 texts (i.e. experiment *B1*) the average gap_{types} is 0.76, while for batches composed of 115 texts (i.e. experiment *B2*) the average gap_{types} is 0.60. Thus, the addition of 40 texts decreased the gap by 0.16 points. However, adding another 35 texts the average gap decreases by 0.27 points (see the gap_{types} for batches with 150 texts). Finally, moving from 150 to 202 texts, the average gap reduces by 0.24 points. Analyzing the previous trend, we argue that at around 200 texts the descent of the curve for the gap_{types} toward value zero slows significantly. Therefore, to further reduce the gap_{types} a large number of texts should be added to the sample corpus.

In particular we notice that by sampling 202 texts, that is the expected value for *n* obtained by applying the proposed statistical method, only 9% of types is lost on average. Practically speaking, the marginal utility related to the number of types in the sample corpus provided by increasing the sample corpus by one text is almost zero. Therefore we can claim that the minimum number of texts *n* suffices for the purposes of collecting a well-sized corpus.

4. Conclusions

A statistical parametric method has been investigated in this paper, to estimate the minimum number of texts to be selected to compile a corpus. Two classical measures of lexical richness (i.e. the TTR and LogTTR) have been used to provide a comparable evaluation index of the texts. The estimate of the variance of the selected indices of lexical richness has been used within the statistical method to provide an accurate estimate of the minimum size of the corpus. Preliminary experiments conducted on texts from a specific domain seem to support the effectiveness of the method to support the construction of a corpus. These introductory but stimulating results encourage us to further study the proposed method and to test other indices, such as the *moving-average type-token ratio* (MATTR), to find a better correlation between the number of texts in the corpus and the relative corpus richness.

Acknowledgments

The authors thank *Prof. Roberto Guarasci* (Università della Calabria, Italy) who has been a springboard for the writing of this paper and *Meik Michalke* from the Heinrich-Heine University (Düsseldorf, Germany) for his precious support in using the *koRpus* package.

References

Baayen R.H. (1996). The effects of lexical specialization on the growth curve of the vocabulary. *Journal of Computational Linguistics*, 22.(4):455-480.

Biber D., Conrad S. and Reppen R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Tognini Bonelli E. and Sinclair J. (2006). Corpora. In Brown K. editor, *Encyclopedia of Language and Linguistics*, 2nd edition. Amsterdam: Elsevier, p. 209.

Broder A., Fontura M., Josifovski V., Kumar R., Motwani R., Nabar S., Panigrahy R., Tomkins A. and Xu, Y. (2006). Estimating Corpus Size via Queries. Proc. of CIKM '06 (15th ACM International Conference on Information and Knowledge Management), pp. 594-603.

Brunet E. (1978). *Le vocabulaire de Jean Giraudoux, structure et évolution*. Genève: Éditions Slatkine.

Chen E.J. (2011). A revisit of two-stage selection procedures. *European Journal of Operational Research*, 210.(2):281–286.

Cortelazzo M. and Tuzzi A. (2008). *Metodi statistici applicati all'italiano*. Bologna: Zanichelli.

Covington M.A. and McFall J.D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17:94-100.

Crane M.A. and Lemoine A.J (1977). *An introduction to the regenerative method for simulation analysis*. In series Lecture Notes in Control and Information Sciences, 4. Berlin, Heidelberg, New York: Springer-Verlag.

De Haan P. (1992). The minimum corpus sample size? In Leitner, G. editor, *New Directions in English Language Corpora: Methodology, Results, Software Development*. New York: Mounton de Grouyter.

Ejiri K. and Smith A.E. (1993). Proposal for a new 'constraint measure' for text. In Köhler R. and Burghard B.R. editors, C*ontributions to Quantitative Linguistics*. Kluwer Academic Publischers, pp. 195–211.

Francis W.N. and Kucera H. (1979). *Brown corpus manual*. Department of Linguistics, Brown University.

Guarasci R. (2008). Indicizzazione e classificazione: concetti generali. In Guarasci R., editor, *Dal documento all'informazione*. Milano:ITER, pp. 215-226.

Guiraud P. (1954). *Les Caractères statistiques du vocabulaire: essais de méthodologie*. Paris:Presses Universitaires de France.

Herdan G. (1960). Type-token mathematics: A Textbook of Mathematical Linguistics. *Journal of the Royal Statistical Society, Series A*, 123.(3):341-342.

Hardie A. and McEnery T. (2006). Statistics. In Brown K. editor, *Encyclopedia of Language and Linguistics*, 2nd edition. Amsterdam: Elsevier, pp. 138-146.

Hillier F.S. and Lieberman G.J. (2010). Supplement 2 to Chapter 20: Regenerative Method. In *Introduction to Operations Research*, 9th edition. McGraw-Hill.

Hunston S. (2006) Corpus Linguistics. In Brown K. editor, *Encyclopedia of Language and Linguistics*, 2nd edition. Amsterdam: Elsevier, pp. 234-248.

Juckett D. (2012). A method for determining the number of documents needed for a gold standard corpus. *Journal of Biomedical Informatics*, 45(3):460-470.

Lauer M. (1995). How much is enough?: Data requirements for statistical NLP. In *Proc. of PACLING'95* (2nd Conference of the Pacific Association for Computational Linguistics), pp. 1-9.

Malvern D. and Richards B. (2012). Measures of Lexical Richness. In Chapelle C.A. editor, *The Encyclopedia of Applied Linguistics*. Malden, MA: Blackwell Wiley, pp. 3622–3627.

Michalke, M. (2013). *koRpus: An R Package for Text Analysis (Version 0.05-1)*. Available from http://reaktanz.de/?c=hacking&s=koRpus

McEnery T. and Wilson A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Pastor G.C. and Domínguez M.S. (2007). Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor. *Procesamiento del Lenguaje Natural*, 39:165-172

Rinott Y. (1978). On two-stage selection procedures and related probability inequalities. *Communications in Statistics - Theory and Methods*, 7.(8):799-811.

Schmidt H. (1994). Improvements in Part-of-Speech Tagging with an Application to German. In *Proc. of ACL SIGDAT-Workshop*. Niemeyer, Tübingen: Feldweg and Hinrichs and Feldweg and Hinrichs editors, pp. 47-50.

Schmidt H. (1995). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. of NeMLaP'95* (International Conference on New Methods in Language Processing), Manchester, UK: Centre for Computational Linguistics, UMIST, pp. 44-49.

Sinclair J. (2004). *Trust the text: language, corpus and discourse*. London: Routledge.

Van Gijsel S., Speelman D. and Geeraerts D. (2006). Locating lexical richness: a corpus linguistic, socio variational analysis. In *Proc. of JADT'06* (8th International Conference on Textual Data Statistical Analysis), 2:961-971.

Woods A., Fletcher P. and Hughes A. (1986). *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Yang D.-H., Gomez P.C. and Song M. (2000). An Algorithm for Predicting the Relation between Lemmas and Corpus Size. *ETRI Journal*, 22.(2):20-31.

Yang D.-H., Lee I.-H. and Cantos P. (2002). On the Corpus Size Needed for Compiling a Comprehensive Computational Lexicon by Automatic Lexical Acquisition. *Computers and the Humanities*, 36.(2):171-190.