A method
for the evaluation of vocabulary overlaps

Dietmar NAJOCK

Résumé : It is shown that the simple percentage of vocabulary which a common reference text covers in various other texts is an insufficient measure of vocabulary overlaps, since it depends heavily on text lengths. A good measure is the expected vocabulary overlap size and its standard deviation—computed on the hypothesis that the texts compared are random extracts from a larger corpus such as the effective vocabulary of an author. Formulae for the overlap size of two and more texts are derived according both to the replacement and the non-replacement model (based on the sequence v_1, v_2, ..., i.e. on the number of lemmata which occur exactly once, twice etc. in the underlying corpus). The formulae are compared with the results of simulated text extraction from the common corpus of *Letters* and *Speeches* of the Greek rhetorician Libanios. The correspondence of formulae and simulation is almost perfect. Of the two models, the non-replacement model is adequate for simple vocabulary sizes, but for overlap sizes the replacement model is better, although a large basic corpus is required. An application to Libanios shows that *Letters* and *Speeches* differ in vocabulary at a very high significance level.

Keywords: Vocabulary statistics, vocabulary overlap, insufficiency of percentages, formula for expected value, formula for standard deviation, random text extraction, replacement model, non-replacement model, simulation, Libanios.

1. An introductory example

Vocabulary statistics are often given in percentages, but do percentages always lead to meaningful results? In a recent study of Cicero's

✉ Priv.-Doz. Dr. Dietmar Najock, Seminar für Klassische Philologie,
Freie Universität Berlin, Ehrenbergstr. 35, D-14195 Berlin.
E-mail: najock@zedat.fu-berlin.de

works on the problem of destination and free will,[1] i.e. on *De Fato*, *De Divinatione* and *De Natura Deorum*, professor Castrillo-Benito compares related works of several later authors: Augustinus' *De Libero Arbitrio*, Lorenzo Valla's *De Libero Arbitrio*, Desiderius Erasmus' *De Libero Arbitrio*, and Martin Luther's *De Servo Arbitrio*. The vocabulary covered by Cicero's 'trilogy' amounts to 51.4 % in the work of Lorenzo Valla, to 40.6 % in Augustinus', to 38.3 % in Erasmus', and to 30.3 % in Luther's work. On account of these values, Castrillo-Benito classifies Valla as the most Ciceronian and Luther as the least Ciceronian author. Comparing the same authors with the *Vulgata*, he finds again Valla's work as most similar (65.5 %) and Luther's as most distinct (52.0 %). Thus he concludes that Luther is the most innovative and Valla the most conservative of the authors compared.[2]

These conclusions, however, are subject to two objections. (i) The vocabulary count by Castrillo-Benito does not refer to lemmata, but to word-forms; this may be misleading in highly inflected languages such as Latin. (ii) The works compared differ considerably in text length: Cicero's 'trilogy' comprises 68,254 word-tokens, Augustinus' work 35,746, Valla's 6,364, Erasmus' 19,540, Luther's 70,224, and the *Vulgata* 661,311.[3] One would suspect that text length has its influence on vocabulary overlaps and that the simple percentage—i.e. the ratio of overlap and complete vocabulary in any text compared—is an insufficient measure: it may be due to the very shortness of Valla's work that the overlap amounts to a high percentage of its vocabulary in both comparisons, and it may be due to the very length of Luther's work that the overlap covers only a low percentage, again in both comparisons. In particular, one cannot hope that the vocabulary of Cicero's 'trilogy' covers much of the *Vulgata*, which is almost ten times as long.

---

[1] CASTRILLO BENITO (Nicolás): 1997, *Racionalismo filosófico y lógica propedéutica en Cicerón: el tratado sobre el destino (De fato)* (Burgos: Servicio de Publicaciones, Universidad de Burgos), pp. 171–175.

[2] The *Vulgata*, with different themes, of course, has only 13.8 % in common with Cicero.

[3] The corresponding type counts are: Cicero 15,200, Augustinus 7,084, Valla 2,408, Erasmus 5,637, Luther 11,822, Vulgata 48,931.

But how can these suspicions be proved or disproved? Little is known on the relation between text lengths and the sizes of vocabulary overlaps. In 1982, L. Ule published a formula for what he calls *vocabulary overlap*,[1] but he defines this overlap by the word-tokens that two texts have in common, not by the word-types.[2] Therefore his formula stands for a different concept and cannot be applied to our problem. In the present paper I derive formulae for both expected values and variances for the size of the vocabulary overlap of two, three or more texts. These formulae are based on two alternative models of random text extraction—with or without replacement of the word-tokens. The second of these models, the non-replacement model, is also the base of Muller's formula for expected vocabulary sizes[3] and of my own formula for the corresponding variance.[4] The common logic of all these formulae will become clear in the following section.

2. The formulae

Let T denote the set of word-tokens of a basic text (or corpus), i.e. the text words in arbitrary order, and W the set of word-types occurring in

---

[1] ULE (Louis): 1982, "Recent Progress in Computer Methods of Authorship Determination", *Association for Literary and Linguistic Computing Bulletin*, X,3, pp. 73–89.

[2] One might prefer, therefore, to call Ule's measure 'expected text overlap'.

[3] MULLER (Charles): 1968, *Initiation à la statistique linguistique* (Paris: Larousse); MULLER (Charles): 1977, *Principes et méthodes de statistique lexicale* (Paris: Hachette). There is continuing interest in the relation of text length n and vocabulary size x: a certain set of equations has recently been compared by PRUSCHA (Helmut): 1998, "Statistical Models for Vocabulary and Text Length with an Application to the NT Corpus", *Literary & Linguistic Computing*, XIII,4 (Oxford: University Press), pp. 195–198. Among the equations tested, the simple relation $x = a \cdot \sqrt{n}$ performs rather well, but the measure proposed for vocabulary richness yields values that are not directly comparable for texts of different lengths.

[4] NAJOCK (Dietmar): 1986, "Bootstrap Experiments for the Evaluation of Expected Values and Variances of Vocabulary Sizes", in BRUNET (Étienne) ed., *Méthodes quantitatives et informatiques dans l'étude des textes* (Genève-Paris: Slatkine-Champion, "Travaux de linguistique quantitative", 35), pp. 657-670. L. Ule used a similar model of random text extraction (without replacement) and recommended that the complete work of an author or some comparable corpus should be used as the base.

this text. Furthermore, let $N = |T|$ denote the number of the word-tokens, i.e. the length of the text, and $V = |W|$ the number of the word-types, i.e. the vocabulary size.

If T_n is a random extract of n word-tokens out of text T and W_n the set of word-types in T_n, then the vocabulary size of the extract is

$$X = |W_n|,$$

a random variable with its own expected value and variance.

Now let $w \in W$ be a certain word-type that occurs exactly i times in T, and let w_o be a certain occurrence of w, i.e. a certain token of this word-type. The probability that w_o—a single occurrence—is not part of the extract T_n may then be estimated by

(1) $p(n) = \mathrm{P}(w_o \notin T_n) = (N-n)/N = 1 - n/N$ in the non-replacement model,

and by

(2) $p(n) = \mathrm{P}(w_o \notin T_n) = ((N-1)/N)^n = (1 - 1/N)^n$ in the replacement model.

The following arguments will invariably use the notation $p(n)$, independent of the model chosen, but in the resulting formulae one of the equations (1) or (2) must be substituted.

The probability that none of the i occurrences of w is part of extract T_n—in as much as the single occurrences may be regarded as independent—is

$$\mathrm{P}(w \notin W_n) = p(n)^i,$$

and the probability that the extract contains (at least once) word-type w is

$$\mathrm{P}(w \in W_n) = 1 - p(n)^i.$$

Similarly,

$$\mathrm{P}(w \in W_m) = 1 - p(m)^i$$

is the probability that another extract, of m word-tokens, contains the same word-type w. If the extracts are independent, the probability that

w occurs in both of them is

$$P(w \in W_n \,\&\&\, w \in W_m) = (1-p(n)^i)\,(1-p(m)^i).$$

Now the vocabulary size is

$$V = |W| = \sum_{w \in W} 1,$$

where the sum is over all word-types of the basic text, and it may also be written as

$$V = \sum_{w \in W} 1 = \sum_i 1 \cdot v_i,$$

where v_i is the number of word-types that occur exactly i times.[1]

Similarly, the expected vocabulary size of the random extract T_n is

$$\begin{aligned} E(X) &= E(|W_n|) \\ &= \sum_{w \in W} 1 \cdot P(w \in W_n) \\ &= \sum_i 1 \cdot v_i \cdot P(w \in W_n \mid w \text{ occurs } i \text{ times}) \\ &= \sum_i 1 \cdot v_i \cdot (1-p(n)^i) \\ &= \sum_i (1-p(n)^i)\, v_i. \end{aligned}$$

In the last expression, which is equivalent to Muller's formula in the non-replacement case, each summand may be regarded as the expected value of a binomially distributed variable with success probability $p = 1 - p(n)^i$ and v_i repetitions of the underlying Bernoulli experiment[2] and, in consequence, with a variance of

$$v_i \cdot p \cdot (1\text{-}p) = (1-p(n)^i)\,(1-(1-p(n)^i))\,v_i = (1-p(n)^i)\,p(n)^i\,v_i.$$

The variance of the vocabulary size of the extract will therefore—further independence presupposed—be

$$\mathrm{Var}(X) = \mathrm{Var}(|W_n|) = \sum_i (1-p(n)^i)\,p(n)^i\,v_i.$$

In the non-replacement case, this corresponds to my formula published in 1986.[3]

---

[1] Ule's formula, too, is based on these v_i and on the extraction lengths involved.

[2] I.e. catching a word of frequency i.

[3] Cf. note 7 above.

Similar considerations may be applied to the vocabulary overlap of two random extracts T_m and T_n. Let Y denote the size of the vocabulary overlap of these extracts. Then the expected value of Y is

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}(|W_m \cap W_n|) \\ &= \sum_{w \in W} 1 \cdot \mathrm{P}(w \in W_m \,\&\&\, w \in W_n) \\ &= \sum_i 1 \cdot v_i \cdot \mathrm{P}(w \in W_m \,\&\&\, w \in W_n \mid w \text{ occurs } i \text{ times}) \\ &= \sum_i 1 \cdot v_i \cdot (1 - p(m)^i)\,(1 - p(n)^i) \\ &= \sum_i (1 - p(m)^i)\,(1 - p(n)^i)\, v_i, \end{aligned}$$

with $p(m)$ and $p(n)$ defined as in (1) or (2) above. Again each summand may be interpreted as the expected value of a binomially distributed variable, so that the size of the vocabulary overlap will have the following variance (independence presupposed):

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(|W_m \cap W_n|) \\ &= \sum_i (1 - p(m)^i)\,(1 - p(n)^i)\,[1 - (1 - p(m)^i)\,(1 - p(n)^i)\,]\, v_i. \end{aligned}$$

These formulae allow comparing the size of an observed vocabulary overlap with its expected value and determining the statistical significance of its deviation, since $Y - \mathrm{E}(Y) \,/\, \sqrt{(\mathrm{Var}(Y)}$ approximately follows the standard normal distribution.[1]

In the same way we may determine expected values and variances for the vocabulary overlap of three or more extracts. The example of three extracts T_l, T_m and T_n will suffice. Let Z denote the size of the vocabulary overlap of these extracts. Then the expected value of Z is

$$\mathrm{E}(Z) = \mathrm{E}(|W_l \cap W_m \cap W_n|) = \sum_i (1 - p(l)^i)\,(1 - p(m)^i)\,(1 - p(n)^i)\, v_i,$$

and the corresponding variance is

$$\begin{aligned} \mathrm{Var}(Z) &= \mathrm{Var}(|W_l \cap W_m \cap W_n|) \\ &= \sum_i (1 - p(l)^i)\,(1 - p(m)^i)\,(1 - p(n)^i)\,[1 - (1 - p(l)^i)\,(1 - \\ &\qquad p(m)^i)\,(1 - p(n)^i)\,]\, v_i. \end{aligned}$$

[1] The distribution of Y is illustrated by Figure 1.

3. Simulation of vocabulary overlaps by bootstrap experiments

3.1 . Design of the experiment

The basic corpus, from which the random texts are drawn, should be large, of course. For the present study I have chosen the combined corpus of the *Letters* and *Speeches* of the Greek rhetorician Libanios, comprising a total of N = 465,744 word-tokens.[1] The vocabulary structure of the corpus is sufficiently described by the sequence v_1, v_2, v_3 etc., i.e. the number of the once occurring words, that of the twice occurring words etc.[2] In our case, this sequence is

2361, 970, 630, 446, 308, ...,

as published in my *Tabulae I.*[3]

Now, since text extraction will be performed by means of random numbers[4] uniformly distributed over the interval [1,N], i.e. over all word-tokens of the basic corpus, these word-tokens may be arbitrarily re-arranged, e.g. the once occurring words may form the beginning, followed by the twice occurring ones etc. Furthermore, each word-type may be replaced by a specific number, since our interest is in word-counts only, not in the words themselves. The basic corpus would then look like this:

1, 2, 3, ..., 2361; 2362, 2362, 2363, 2363, ..., 3331, 3331; 3332, 3332, 3332, ...

In this sequence, the type-number associated with a random token-number can be determined by a suitable algorithm.[5] The word-type-

---

[1] And V = 8,011 word-types.

[2] Note that $v_1 + v_2 + v_3 + ... = V$ and $1 \cdot v_1 + 2 \cdot v_2 + 3 \cdot v_3 + ... = N$.

[3] NAJOCK (Dietmar): 1996, *Concordantiae in Libanium, Pars Quarta, Tabulae*, I (Hildesheim: Olms-Weidmann), pp. 869–871.

[4] The random numbers are generated by *drand48*, a C-function. The Fortran-function *rand* proved to be insufficient.

[5] Each type-frequency i defines a 'passage'. Once the passage is found by binary search, the type-number is easily computed.

number serves as an index for a frequency list of the word-types drawn. Each random number increases the corresponding entry in this frequency list by 1, so the program keeps trace of the vocabulary growth[1] of an individual sample text, say extract-1. At certain text lengths of extract-1, e.g. at multiples of 5,000, the same procedure is performed for a second, independent sample, say extract-2, with its own frequency list and with stops at the same text lengths. At these stops, i.e. at particular combinations of two text lengths, the size of the vocabulary overlap of the extracts as well as the simple vocabulary size of extract-1 are stored in appropriate arrays. The whole procedure described is repeated 1,000 times, so that, if 10 stages of text length growth are regarded, a 1,000 by 10 by 10 array is needed to store the vocabulary overlap sizes. This array allows computing, for each of the text length combinations, a pseudo-empirical mean and a pseudo-empirical variance (or standard deviation). Both of these values are calculated from 1,000 instances, and the results are included in 10 by 10 tables as given below.

Since each of the random numbers used can occur more than once, the sampling performed is 'with replacement' of the word-tokens. This also means that the sequence v_1, v_2, v_3, ... defines a probability distribution of the word-types: a word-type occurring i times is drawn with probability i/N. Furthermore, an extract may grow to arbitrary length in this model, but its vocabulary cannot exceed that of the basic corpus. Therefore the vocabulary sizes, especially of long extracts, will be lower than the real ones—unless the basic corpus comprises the complete effective vocabulary that could have been used in the texts regarded. Similar considerations must be made for the vocabulary overlap.[2]

The formulae based on extraction without replacement will, of course, show a different behaviour. When the extracts become as long as the basic corpus, the complete basic vocabulary will be exhausted; the overlap will also be exhaustive, and both variances decline to zero.

---

[1] And, of course, of the text length growth.

[2] Cf. Section 3.3 below.

Longer extracts are not provided by the non-replacement model.[1]

Extraction with replacement appears to be the more realistic model, not only because it allows text extracts of arbitrary length: real text production seems to be performed by repeated recurrence to the same unabridged set of word-types as given by a specific *lexique* or effective vocabulary, with various probabilities for the single word-types. For practical purposes, such a *lexique* can only be approached by a very large corpus. Libanios' *Letters*, e.g., comprise 215,354 word-tokens and 5,149 word-types, but I have shown that at least 6,250 types should be added within this genre.[2] So even the combined corpus of *Letters* and *Speeches*, with 8,011 word-types, is still too small in comparison with the presumable effective vocabulary of Libanios.

3.2. Results

Results are given in groups of tables, each group comprising separate tables for tiny, small, medium and large extracts. Tiny extracts range from 50 to 500 word-tokens, small ones from 500 to 5,000, medium ones from 5,000 to 50,000, and large ones from 50,000 to 500,000. The lengths given in the table headlines refer to extract-2, the lengths on the left margin to extract-1. Most tables are symmetric with respect to extract-1 and extract-2, but the first type of tables to be described here is not.

3.2.1. *Tables 1.1 to 1.4*

The first set of tables displays the percentage of vocabulary that a common reference text—here extract-2—covers in any compared

---

[1] In spite of these limitations, Muller's formula for the expected vocabulary size gives reasonable results for sample texts with lengths up to $2N$, and my formula for the corresponding variance may reasonably be used up to text length $N/2$.

[2] This result will be published in my *Sprachstatistische Untersuchungen zu den Briefen und Reden des Libanios*. It follows from the fact that the best mathematical model for the sequence $v_1, v_2, v_3, \ldots$ provides also an estimation for v_0, the number of unused words of the effective vocabulary. The best model is the hypergeometric one as described by DOLPHIN (Bernard): 1979, *Vocabulaire et lexique* (Genève: Slatkine, "Travaux de linguistique quantitative", 7), pp. 84–85.

text—here extract-1. A second row (marked by V) contains the absolute vocabulary size to which the percentage must be referred—that of extract-1. All table entries result from the simulation described. These tables show that the mean vocabulary overlap, expressed as percentage of any compared vocabulary, depends heavily on the text lengths (or extract lengths) involved: short reference texts will not cover as much as long ones, and in the compared texts, the percentage covered decreases with increasing text length. Therefore, a mere comparison of such percentages may be highly misleading. In fact, the percentages vary between 8.5 % and 96.8 % in our examples, depending only on the text lengths involved.

3.2.2. *Tables 2.1 to 2.4*

The second set of tables displays the mean vocabulary overlap as given by simulation (S) and the expected vocabulary overlap according to the formulae (F), both in the replacement model (Fr) and in the non-replacement model (Fn). Thus there are three rows for each length of extract-1, marked by S, Fr and Fn.

(a) The values resulting from the formula of the replacement model (Fr) stay very close to the simulation values over all extraction lengths. Indeed they should do so, since the simulation procedure, too, works with replacement of the word-tokens. The good fit shows that the formula does not suffer from any computational bias.

(b) The formula of the non-replacement model, however, yields values that—albeit close to the simulation values for shorter extracts—exceed them more and more, as extraction length increases. Indeed the non-replacement model comes close to the replacement model for short extracts, and both models are equal for extraction length 1. On the other hand, extraction without replacement becomes exhaustive, when the length of the basic text or corpus is reached, in our case with 465,744 word-tokens. The vocabulary overlap of two extracts of this length must be complete, i.e. it should equal 8,011 word-types. In fact the non-replacement formula yields almost exactly this value: linear interpolation between 7,851.8 and 8,361.1, the expected values for the overlaps of two extracts of 450,000 and two extracts of 500,000 words

respectively, yields 7,851.8 + 160.4 = 8,012.2. Again we can state that the formula agrees well with the underlying model and that the computation may be regarded as free from errors.

3.2.3. *Tables 3.1 to 3.8*

The third set of tables shows the difference of the expected values (Fr, Fs) and the mean vocabulary overlap (S)—as given in the previous set of tables—expressed in percentages of the latter.[1] These tables confirm, in a more simple form, the results mentioned above. In particular they show that the formula for the non-replacement model and our simulation with replacement fit well together—with deviations less than 1 %—if *at least one* of the extracts does not exceed extraction length 5,000. In this case, the second extract may well reach text length 50,000. The difference may still be regarded as tolerable—with deviations less than ca. 6 %—as long as *at least one* of the extracts does not exceed extraction length 50,000, i.e. ca. *N*/10. The other extract may well reach length *N* (= 465,744) or 500,000 in this case. If both extracts reach length *N*, however, the difference to simulation with replacement is ca. 30 %.

3.2.4. *Tables 4.1 to 4.4*

This set of tables shows the pseudo-empirical standard deviation of the vocabulary overlap according to our simulation (S) and the theoretical standard deviation according to the formulae, both that of the replacement model (Fr) and that of the non-replacement model (Fn). Again there are three rows for each length of extract-1, marked by S, Fr and Fn respectively. All three values (S, Fr and Fn) remain close together, if *at least one* of the extracts is not much longer than 250,000 word-tokens, i.e. ca. *N*/2.[2] For short extracts, both formulae yield almost identical values, very close to simulation, but always a little higher. For medium extract lengths, and even for extract lengths up to ca.

[1] This may be denoted by 100·(Fr – S)/S or 100·(Fn – S)/S.

[2] Note that this is also the upper bound for a good fit of the standard deviation of the simple vocabulary size.

250,000, i.e. ca. *N*/2, the values of the non-replacement formula are slightly above those of the replacement formula and those of simulation, but for longer extracts they decrease more and more, below both the replacement formula values and the simulation values. At length 450,000 (of both extracts) it is less than half the other values (12.2 < 28.9 = 28.9), and if at least one of the extracts has length 500,000, it is so close to zero that, on account of underflow, the output gives *nan*, i.e. *not a number*. Indeed at length 465,744 both extracts must be exhaustive in the non-replacement model, and the vocabulary overlap cannot vary: variance and standard deviation must be zero. All this means that the non-replacement formula for the standard deviation corresponds well to the underlying model, without computational bias. The same holds for the replacement formula, since it keeps close to simulation throughout.[1] It should be noted, however, that the replacement formula yields its maximum (29.6) at extraction length 300,000 (of both extracts), with a gradual decline thereafter.[2] Thus the limitations of both models become apparent at about extract length *N*/2, but they remain moderate in the replacement model.

3.2.5. *Tables 5.1 to 5.8*

These tables show the difference of the theoretical standard deviations (Fr, Fn) and the pseudo-empirical standard deviation (S)—as given in the previous set of tables—expressed in percentages of the latter, i.e. of the simulation values.

(a) The replacement model: For long extracts, the differences vary around 0 %, but the positive differences prevail (about two thirds). In the medium range, there are only three negative differences, and there are none in the range of short extracts. The shorter the extracts the higher the differences, especially when they are measured in percentages. Although the absolute differences are often less than 1.0 and only very seldom greater than 2.0, such values yield rather high percentages, when referred to the small standard deviations of short texts,

[1] Only for very short extracts a certain difference must be stated, see 3.2.5 below.

[2] Simulation behaves similarly.

in particular if one of the extracts is shorter than 1,000 word-tokens. The reason for this deviation from a perfect fit might be some mysterious property of the random number generator or a computational bias resulting from an accumulation of rounding errors.[1] In the latter case, a correction could be obtained by subtracting a number between 1.0 and 0.0 from the standard deviation, depending on the length of the extract.[2]

(b) The non-replacement formula behaves similar, except for the tendency described above: for extracts longer than ca. *N*/2 the standard deviation declines more and more and finally approaches zero.

3.3. Discussion of the results

Up to now we have shown that the formulae correspond well with the models, except perhaps the standard deviation for very short extracts. But which of the models corresponds better to reality? The previous remarks on real text production would favour extraction with replacement, but there remains the problem that only seldom a given basic text or text corpus will be sufficiently representative of an author's effective vocabulary (especially if lemmatization is requested). Therefore in many cases, as in the Libanios example, at least the vocabulary sizes of long extracts will be too low, if computed according to the replacement model. The non-replacement model, however, yields just the vocabulary size of the basic text, when an extract reaches length *N* and becomes exhaustive. This means that the non-replacement model, in as much as the basic corpus may be regarded as typical, gives exactly the vocabulary sizes that should be expected. Obviously this model is the superior choice with respect to the simple vocabulary size and its variance.[3]

As to vocabulary overlaps, the case is slightly more complicated. Consider two random extracts of fixed length, and for simplicity

---

[1] But note that my Fortran-program works with double precision numbers in this part.

[2] 1.0 for short (or tiny) extracts, 0.0 for extracts of medium length, with intermediate values.

[3] Within the known limits, of course.

assume that both are equal in length. If they are drawn from a much larger basic corpus, there is much room for variation in vocabulary so that the overlap will be relatively small. One might call this effect a *tendency towards vocabulary diversity* in large basic corpora.[1] On the other hand, a small basic corpus has a small vocabulary and thereby sets an inevitable limit to any vocabulary overlap. Thus there is a second effect that yields small overlaps: the *tendency towards vocabulary constriction* in small corpora. Since both large and small corpora give rise to small overlaps, only a corpus of medium size will yield larger ones. The largest overlaps seem to result from corpora of about double extract length or from slightly longer ones, as will be seen in the Libanios example. The same example will show that much longer corpora yield smaller overlaps, but that the decrease in overlap size is rather moderate in this case.

Unfortunately, it has not been possible to derive from the given sequence v_1, v_2, v_3, ...—with $N = 465{,}744$ and $V = 8{,}011$ in the case of Libanios—a corresponding sequence v_1', v_2', v_3', ... for a hypothetical longer corpus, say of length $2N$. But the above mentioned effects can also be demonstrated by comparing the results derived from artificially shortened corpora. For that purpose I extracted random partial corpora[2] of the lengths 50,000, 100,000 etc. up to 450,000, more exactly five versions for each of these lengths (N'). Thus I got five more or less similar sequences v_1', v_2', v_3', ... for each length and five more or less similar vocabulary sizes V'. While the five versions of the sequence v_1', v_2', v_3', ... could not be transformed into something like a 'mean sequence', the resulting vocabulary overlap sizes and the vocabulary sizes V' could be used to compute means. The mean overlap sizes are given in Table 6, but the rightmost column, which refers to the complete given corpus, does not contain means. Each row of the table refers to a pair of extracts of equal length (as given in the leftmost column).

---

[1] Similarly, small extracts yield relatively small overlaps.

[2] These extractions must be without replacement, since the results should be comparable with those of the complete given corpus (this corpus may be regarded as an exhaustive non-replacement extract from itself).

The reader will notice that Table 6 confirms all the above statements. The table also shows that even a relatively large corpus like that of the Libanios Letters and Speeches yields overlaps that are still too large: if the effective vocabulary of Libanios were known (so that the table could be extended to the right side), the values would be somewhat lower than in the rightmost column. The non-replacement model, however, gives higher values so that it must be regarded as inadequate for vocabulary overlaps. The replacement model is more convenient, but obviously a very large basic corpus is required, a corpus that should cover the effective vocabulary of the author or the authors regarded as closely as possible. Minor deviations from the effective vocabulary, especially towards a larger one, seem to be tolerable.

4. An application to Libanios

Can the *Letters* and *Speeches* of Libanios be regarded as drawn at random from a common effective vocabulary? The relevant data are the following:

	word-tokens	word-types
Letters	215,354	5,149
Speeches	250,390	6,974
The combined corpus	465,744	8,011

The actual vocabulary overlap of *Letters* and *Speeches* comprises 5,149 + 6,974 – 8,011 = 4,112 word-types.[1] The expected vocabulary overlap must be computed by linear interpolation from the next entries in Table 2.4. With the *Letters* as extract-1 (rows) and the *Speeches* as extract-2 (columns), these entries are as follows:

[1] Thus the word types occurring only in the *Speeches* are 6,974 – 4,112 = 2,862, while those occurring in the *Letters* only are as few as 5,149 – 4,112 = 1,037, i.e. the specific vocabulary of the *Speeches* is much larger.

	250,000	*250,390*	300,000
200,000	4,884.5		5,028.4
215,354	$x_1 = ?$	y = ??	$x_2 = ?$
250,000	5,101.5		5,264.4

In a first step we find $x_1 = 4{,}951.1$ and $x_2 = 5{,}100.9$, in a second the expected vocabulary overlap $y = 4{,}952.3$. This value is not far from 5,101.5 (the value for two extracts of length 250,000), and Table 6 as well as Figure 2 show that it would almost certainly not fall below 4,800, if it could be based on a larger corpus.[1] Thus the expected vocabulary overlap exceeds the actual one by about 700 word-types or 23 standard deviations.[2] Comparing the standard normal distribution, we find

$$z \leq (4{,}112 - 4{,}800) / 29.4 = -23.4,$$

i.e. the small size of the actual vocabulary overlap is extremely significant. We may conclude that the vocabulary of the *Letters* and *Speeches* cannot be regarded as drawn at random from a common pool. On the contrary, we must assume that the *Letters* and *Speeches* are decidedly different with respect to word usage.[3]

5. Conclusion

There is some hope that the statistical study of Cicero's works on destination and free will and of related later works can be repeated with lemmatized data. The combined corpus of the *Vulgata* and a major portion of Cicero seems to constitute a suitable base for this purpose.

---

[1] I.e. a corpus that might be regarded as fully representative for Libanios' vocabulary.

[2] According to Table 4.4, the standard deviation is about 29.4 for texts of these lengths.

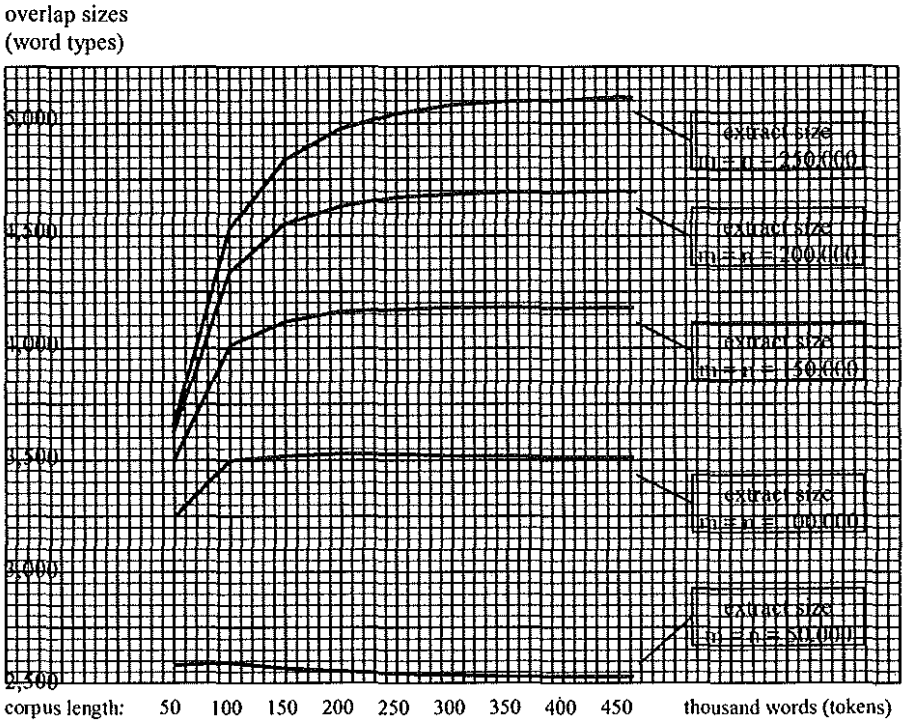
[3] In fact I have shown some of the differences in my *Sprachstatistische Untersuchungen zu den Briefen und Reden des Libanios* (to be published in the near future). They refer mainly to special topics of both literary genres and to a certain avoidance of seldom words in the *Letters*. The vocabulary is much richer in the *Speeches*, especially in the epideictic ones.

Figure 1:
Distribution of vocabulary overlap sizes for text lengths 2000 and 2000 [next page]

Lower bound: 254 , upper bound: 321, mean: 286.6.
Variance: 102.336, standard deviation: 10.116, skewness: 0.078.
0.5 %-tails (or less) cut off in diagram.

```
          0         1         2         3         4         5         6
          01234567890123456789012345678901234567890123456789012345678901234567890
  1   262 I*
  6   263 I******
  3   264 I***
  2   265 I**
  6   266 I******
  6   267 I******
 11   268 I***********
  4   269 I****
 11   270 I***********
  7   271 I*******
 13   272 I*************
 11   273 I***********
 16   274 I****************
 24   275 I************************
 30   276 I******************************
 25   277 I************************
 30   278 I******************************
 32   279 I********************************
 36   280 I************************************
 26   281 I**************************
 29   282 I*****************************
 47   283 I***********************************************
 46   284 I**********************************************
 38   285 I**************************************
 45   286 I*********************************************
 42   287 I*******************************************
 38   288 I***************************************
 37   289 I**************************************
 31   290 I*******************************
 35   291 I***********************************
 37   292 I**************************************
 30   293 I******************************
 24   294 I************************
 30   295 I******************************
 17   296 I*****************
 19   297 I*******************
 19   298 I*******************
 20   299 I********************
 26   300 I**************************
 13   301 I*************
 16   302 I****************
  5   303 I*****
 11   304 I***********
  7   305 I*******
  5   306 I*****
  2   307 I**
  6   308 I******
  2   309 I**
  8   310 I********
  5   311 I*****
          0         1         2         3         4         5         6
          01234567890123456789012345678901234567890123456789012345678901234567890
```

Figure 2:
Mean vocabulary overlap, based on corpora of lengths 50,000 to 450,000



Tables 1.1 - 1.2:
Vocabulary overlap in percentage of vocabulary size of extract-1

	50	100	150	200	250	300	350	400	450	500
50 %	20.5	29.5	35.5	39.8	43.2	46.1	48.5	50.6	52.4	54.0
50 V	38	38	38	38	38	38	38	38	38	38
100 %	16.5	24.9	30.6	35.0	38.5	41.4	43.8	45.9	47.7	49.3
100 V	69	69	69	69	69	69	69	69	69	69
150 %	14.2	22.0	27.5	31.7	35.3	38.1	40.6	42.8	44.7	46.4
150 V	96	96	96	96	96	96	96	96	96	96
200 %	12.7	20.1	25.4	29.5	32.9	35.7	38.1	40.2	42.1	43.8
200 V	120	120	120	120	120	120	120	120	120	120
250 %	11.7	18.6	23.6	27.6	30.9	33.7	36.2	38.3	40.2	41.9
250 V	143	143	143	143	143	143	143	143	143	143
300 %	10.7	17.4	22.2	26.1	29.3	32.1	34.5	36.5	38.5	40.2
300 V	165	165	165	165	165	165	165	165	165	165
350 %	10.0	16.3	21.0	24.8	28.0	30.7	33.0	35.2	37.0	38.7
350 V	185	185	185	185	185	185	185	185	185	185
400 %	9.5	15.4	19.9	23.6	26.7	29.4	31.7	33.8	35.6	37.3
400 V	205	205	205	205	205	205	205	205	205	205
450 %	8.9	14.6	19.0	22.6	25.7	28.3	30.6	32.7	34.5	36.2
450 V	224	224	224	224	224	224	224	224	224	224
500 %	8.5	14.0	18.2	21.8	24.8	27.4	29.6	31.7	33.5	35.2
500 V	242	242	242	242	242	242	242	242	242	242

	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
500 %	35.2	46.8	53.7	58.5	62.2	65.1	67.5	69.6	71.2	72.8
500 V	243	243	243	243	243	243	243	243	243	243
1000 %	28.4	39.7	46.8	52.0	55.9	59.2	61.9	64.1	66.1	67.9
1000 V	400	400	400	400	400	400	400	400	400	400
1500 %	24.6	35.3	42.4	47.6	51.8	55.1	58.0	60.5	62.6	64.4
1500 V	529	529	529	529	529	529	529	529	529	529
2000 %	22.2	32.4	39.4	44.6	48.8	52.2	55.1	57.6	59.8	61.8
2000 V	642	642	642	642	642	642	642	642	642	642
2500 %	20.3	30.1	36.9	42.1	46.3	49.8	52.7	55.3	57.5	59.5
2500 V	743	743	743	743	743	743	743	743	743	743
3000 %	19.0	28.4	35.0	40.2	44.4	47.8	50.8	53.4	55.7	57.7
3000 V	834	834	834	834	834	834	834	834	834	834
3500 %	17.8	26.9	33.4	38.5	42.7	46.1	49.1	51.7	54.0	56.1
3500 V	919	919	919	919	919	919	919	919	919	919
4000 %	16.9	25.7	32.1	37.1	41.2	44.6	47.6	50.3	52.6	54.7
4000 V	998	998	998	998	998	998	998	998	998	998
4500 %	16.2	24.7	30.9	35.8	39.9	43.3	46.3	48.9	51.2	53.3
4500 V	1071	1071	1071	1071	1071	1071	1071	1071	1071	1071
5000 %	15.5	23.8	29.9	34.7	38.7	42.1	45.1	47.7	50.0	52.1
5000 V	1141	1141	1141	1141	1141	1141	1141	1141	1141	1141

Tables 1.3 - 1.4:
Vocabulary overlap in percentage of vocabulary size of extract-1

	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
5000 %	52.2	65.6	72.6	77.1	80.1	82.5	84.3	85.7	86.9	87.9
5000 V	1139	1139	1139	1139	1139	1139	1139	1139	1139	1139
10000 %	44.4	58.2	66.0	71.0	74.7	77.5	79.7	81.5	83.0	84.3
10000 V	1686	1686	1686	1686	1686	1686	1686	1686	1686	1686
15000 %	39.9	53.5	61.5	66.9	70.9	74.0	76.4	78.4	80.1	81.5
15000 V	2080	2080	2080	2080	2080	2080	2080	2080	2080	2080
20000 %	36.8	50.1	58.2	63.7	67.9	71.1	73.7	75.8	77.6	79.2
20000 V	2393	2393	2393	2393	2393	2393	2393	2393	2393	2393
25000 %	34.5	47.5	55.5	61.2	65.4	68.8	71.5	73.8	75.7	77.3
25000 V	2654	2654	2654	2654	2654	2654	2654	2654	2654	2654
30000 %	32.7	45.4	53.4	59.1	63.4	66.8	69.6	71.9	73.9	75.7
30000 V	2880	2880	2880	2880	2880	2880	2880	2880	2880	2880
35000 %	31.2	43.6	51.5	57.2	61.6	65.1	67.9	70.3	72.4	74.2
35000 V	3080	3080	3080	3080	3080	3080	3080	3080	3080	3080
40000 %	30.0	42.2	50.0	55.7	60.1	63.6	66.5	68.9	71.1	72.9
40000 V	3259	3259	3259	3259	3259	3259	3259	3259	3259	3259
45000 %	29.0	40.9	48.7	54.3	58.7	62.3	65.2	67.7	69.8	71.7
45000 V	3422	3422	3422	3422	3422	3422	3422	3422	3422	3422
50000 %	28.1	39.8	47.5	53.1	57.5	61.0	64.0	66.5	68.7	70.6
50000 V	3571	3571	3571	3571	3571	3571	3571	3571	3571	3571

	50000	100000	150000	200000	250000	300000	350000	400000	450000	500000
50000 %	70.6	81.9	87.1	90.2	92.3	93.7	94.8	95.6	96.3	96.8
50000 V	3569	3569	3569	3569	3569	3569	3569	3569	3569	3569
100000 %	63.3	76.0	82.5	86.4	89.2	91.1	92.6	93.7	94.7	95.4
100000 V	4619	4619	4619	4619	4619	4619	4619	4619	4619	4619
150000 %	59.0	72.3	79.4	83.8	86.9	89.2	91.0	92.3	93.4	94.3
150000 V	5270	5270	5270	5270	5270	5270	5270	5270	5270	5270
200000 %	56.2	69.7	77.1	81.9	85.3	87.8	89.7	91.2	92.5	93.5
200000 V	5730	5730	5730	5730	5730	5730	5730	5730	5730	5730
250000 %	54.2	67.7	75.3	80.3	83.9	86.6	88.7	90.3	91.7	92.8
250000 V	6080	6080	6080	6080	6080	6080	6080	6080	6080	6080
300000 %	52.6	66.2	74.0	79.1	82.8	85.6	87.8	89.6	91.0	92.2
300000 V	6357	6357	6357	6357	6357	6357	6357	6357	6357	6357
350000 %	51.4	65.0	72.8	78.1	81.9	84.8	87.1	88.9	90.4	91.7
350000 V	6582	6582	6582	6582	6582	6582	6582	6582	6582	6582
400000 %	50.5	64.0	71.9	77.3	81.2	84.1	86.5	88.4	90.0	91.3
400000 V	6768	6768	6768	6768	6768	6768	6768	6768	6768	6768
450000 %	49.7	63.2	71.1	76.5	80.5	83.6	86.0	87.9	89.6	90.9
450000 V	6924	6924	6924	6924	6924	6924	6924	6924	6924	6924
500000 %	49.0	62.5	70.5	75.9	80.0	83.1	85.5	87.5	89.2	90.6
500000 V	7058	7058	7058	7058	7058	7058	7058	7058	7058	7058

Tables 2.1 - 2.2: Mean vocabulary overlap according to simulation (S) and expected value according to formulae (Fr and Fn)

	50	100	150	200	250	300	350	400	450	500
50 S	7.9	11.3	13.6	15.3	16.6	17.7	18.6	19.4	20.1	20.7
50 Fr	7.7	11.3	13.6	15.3	16.6	17.7	18.6	19.3	20.0	20.6
50 Fn	7.7	11.3	13.6	15.3	16.6	17.7	18.6	19.4	20.0	20.6
100 S	11.4	17.2	21.1	24.1	26.6	28.5	30.2	31.7	32.9	34.0
100 Fr	11.3	17.1	21.1	24.1	26.5	28.5	30.2	31.6	32.9	34.0
100 Fn	11.3	17.1	21.1	24.1	26.5	28.5	30.2	31.6	32.9	34.0
150 S	13.7	21.2	26.4	30.5	33.9	36.6	39.0	41.1	43.0	44.6
150 Fr	13.6	21.1	26.5	30.6	33.9	36.6	39.0	41.0	42.8	44.4
150 Fn	13.6	21.1	26.5	30.6	33.9	36.6	39.0	41.0	42.8	44.4
200 S	15.3	24.3	30.6	35.6	39.7	43.1	46.0	48.5	50.7	52.9
200 Fr	15.3	24.1	30.6	35.6	39.6	43.1	46.0	48.6	50.9	52.9
200 Fn	15.3	24.1	30.6	35.6	39.6	43.1	46.0	48.6	50.9	52.9
250 S	16.8	26.7	33.9	39.6	44.4	48.4	51.9	55.0	57.8	60.2
250 Fr	16.6	26.5	33.9	39.6	44.4	48.4	51.9	54.9	57.7	60.1
250 Fn	16.6	26.5	33.9	39.6	44.4	48.4	51.9	55.0	57.7	60.1
300 S	17.7	28.7	36.7	43.0	48.4	53.1	56.9	60.3	63.5	66.3
300 Fr	17.7	28.5	36.6	43.1	48.4	53.0	56.9	60.4	63.5	66.4
300 Fn	17.7	28.5	36.6	43.1	48.4	53.0	56.9	60.4	63.6	66.4
350 S	18.6	30.2	39.0	46.0	51.9	57.0	61.4	65.3	68.8	71.9
350 Fr	18.6	30.2	39.0	46.0	51.9	56.9	61.3	65.2	68.7	71.9
350 Fn	18.6	30.2	39.0	46.0	51.9	56.9	61.3	65.2	68.7	71.9
400 S	19.5	31.7	41.0	48.5	54.9	60.4	65.2	69.4	73.2	76.7
400 Fr	19.3	31.6	41.0	48.6	54.9	60.4	65.2	69.5	73.3	76.8
400 Fn	19.4	31.6	41.0	48.6	55.0	60.4	65.2	69.5	73.4	76.9
450 S	19.9	32.8	42.6	50.7	57.7	63.5	68.7	73.3	77.4	81.2
450 Fr	20.0	32.9	42.8	50.9	57.7	63.5	68.7	73.3	77.5	81.3
450 Fn	20.0	32.9	42.8	50.9	57.7	63.6	68.7	73.4	77.5	81.3
500 S	20.7	34.0	44.3	52.9	60.2	66.5	71.9	76.9	81.4	85.5
500 Fr	20.6	34.0	44.4	52.9	60.1	66.4	71.9	76.8	81.3	85.4
500 Fn	20.6	34.0	44.4	52.9	60.1	66.4	71.9	76.9	81.3	85.4

	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
500 S	85.5	113.6	130.4	142.1	151.1	158.2	164.1	169.1	173.2	176.9
500 Fr	85.4	113.5	130.3	142.1	151.0	158.1	164.0	168.9	173.2	176.9
500 Fn	85.4	113.5	130.4	142.2	151.2	158.3	164.2	169.1	173.4	177.1
1000 S	113.7	158.8	187.5	208.1	223.9	236.9	247.7	256.8	264.7	271.8
1000 Fr	113.5	158.4	187.0	207.7	223.8	236.8	247.6	256.7	264.7	271.6
1000 Fn	113.5	158.5	187.2	208.0	224.1	237.1	248.0	257.2	265.1	272.1
1500 S	130.4	187.1	224.4	252.2	274.3	292.0	307.2	320.3	331.6	341.2
1500 Fr	130.3	187.0	224.6	252.4	274.3	292.2	307.3	320.1	331.3	341.1
1500 Fn	130.4	187.2	224.9	252.8	274.8	292.8	307.9	320.8	332.0	341.9
2000 S	142.4	207.9	252.8	286.6	313.5	335.4	354.1	370.1	384.3	396.7
2000 Fr	142.1	207.7	252.4	286.1	313.0	335.1	353.8	369.9	384.0	396.5
2000 Fn	142.2	208.0	252.8	286.6	313.6	335.8	354.6	370.8	385.0	397.5
2500 S	150.7	223.6	274.0	312.8	344.0	369.8	391.8	410.7	427.3	442.1
2500 Fr	151.0	223.8	274.3	313.0	344.0	369.9	391.8	410.9	427.5	442.3
2500 Fn	151.2	224.1	274.8	313.6	344.8	370.8	392.8	412.0	428.8	443.6
3000 S	158.4	237.0	292.4	335.4	370.2	399.1	424.0	445.7	464.7	481.5
3000 Fr	158.1	236.8	292.2	335.1	369.9	398.9	423.8	445.4	464.5	481.4
3000 Fn	158.3	237.1	292.8	335.8	370.8	400.0	425.0	446.7	465.9	482.9
3500 S	163.8	247.6	307.3	354.0	392.2	424.0	451.3	475.3	496.6	515.7
3500 Fr	164.0	247.6	307.3	353.8	391.8	423.8	451.3	475.2	496.4	515.3
3500 Fn	164.2	248.0	307.9	354.6	392.8	425.0	452.6	476.8	498.1	517.1
4000 S	168.5	256.4	320.0	370.0	410.8	445.5	475.4	501.6	524.9	545.7
4000 Fr	168.9	256.7	320.1	369.9	410.9	445.4	475.2	501.3	524.4	545.1
4000 Fn	169.1	257.2	320.8	370.8	412.0	446.7	476.8	503.0	526.3	547.1
4500 S	173.3	264.8	331.3	383.9	427.3	464.1	496.2	524.5	549.2	571.6
4500 Fr	173.2	264.7	331.3	384.0	427.5	464.5	496.4	524.4	549.3	571.6
4500 Fn	173.4	265.1	332.0	385.0	428.8	465.9	498.1	526.3	551.4	573.9
5000 S	176.7	271.4	340.9	396.2	441.9	481.0	515.0	544.8	571.4	595.0
5000 Fr	176.9	271.6	341.1	396.5	442.3	481.4	515.3	545.1	571.6	595.5
5000 Fn	177.1	272.1	341.9	397.5	443.6	482.9	517.1	547.1	573.9	597.9

Tables 2.3 - 2.4: Mean vocabulary overlap according to simulation (S) and expected value according to formulae (Fr and Fn)

	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
5000 S	595.3	747.8	827.6	878.1	913.3	939.6	960.3	976.8	990.7	1002.2
5000 Fr	595.5	748.2	828.6	879.4	914.9	941.3	961.9	978.5	992.2	1003.7
5000 Fn	597.9	752.3	833.7	885.3	921.5	948.6	969.6	986.6	1000.7	1012.5
10000 S	748.5	981.4	1112.6	1198.4	1260.6	1307.4	1344.7	1375.1	1400.2	1421.4
10000 Fr	748.2	981.3	1112.4	1198.8	1260.7	1307.8	1345.0	1375.3	1400.4	1421.8
10000 Fn	752.3	988.5	1122.0	1210.2	1273.8	1322.1	1360.5	1391.7	1417.8	1439.9
15000 S	829.4	1112.6	1279.4	1392.2	1475.2	1538.8	1590.0	1631.6	1666.5	1696.1
15000 Fr	828.6	1112.4	1279.1	1391.8	1474.4	1538.1	1589.0	1630.7	1665.7	1695.6
15000 Fn	833.7	1122.0	1292.2	1407.9	1493.0	1558.7	1611.4	1654.8	1691.3	1722.5
20000 S	879.5	1199.2	1392.0	1524.9	1624.2	1701.0	1763.1	1814.6	1858.1	1895.3
20000 Fr	879.4	1198.8	1391.8	1525.2	1624.4	1701.8	1764.1	1815.7	1859.3	1896.6
20000 Fn	885.3	1210.2	1407.9	1545.2	1647.7	1727.9	1792.8	1846.7	1892.3	1931.5
25000 S	915.3	1259.7	1473.8	1623.8	1736.6	1825.3	1897.8	1958.2	2008.8	2052.7
25000 Fr	914.9	1260.7	1474.4	1624.4	1737.3	1826.1	1898.3	1958.4	2009.4	2053.3
25000 Fn	921.5	1273.8	1493.0	1647.7	1764.7	1857.2	1932.7	1995.7	2049.4	2095.8
30000 S	941.1	1308.2	1537.3	1701.6	1825.8	1924.2	2004.9	2072.5	2130.2	2180.0
30000 Fr	941.3	1307.8	1538.1	1701.8	1826.1	1924.8	2005.5	2073.1	2130.7	2180.4
30000 Fn	948.6	1322.1	1558.7	1727.9	1857.2	1960.2	2044.9	2116.1	2177.0	2229.8
35000 S	960.7	1343.2	1586.9	1762.0	1896.5	2003.8	2092.2	2166.9	2230.3	2285.3
35000 Fr	961.9	1345.0	1589.0	1764.1	1898.3	2005.5	2093.7	2167.8	2231.3	2286.4
35000 Fn	969.6	1360.5	1611.4	1792.8	1932.7	2044.9	2137.7	2216.1	2283.5	2342.1
40000 S	979.2	1376.0	1630.8	1815.3	1957.9	2072.3	2166.5	2246.8	2315.7	2375.7
40000 Fr	978.5	1375.3	1630.7	1815.7	1958.4	2073.1	2167.8	2247.9	2316.6	2376.4
40000 Fn	986.6	1391.7	1654.8	1846.7	1995.7	2116.1	2216.1	2301.0	2374.2	2438.2
45000 S	992.1	1400.9	1666.2	1859.9	2009.9	2130.9	2231.4	2316.8	2390.5	2454.5
45000 Fr	992.2	1400.4	1665.7	1859.3	2009.4	2130.7	2231.3	2316.6	2390.1	2454.3
45000 Fn	1000.7	1417.8	1691.3	1892.3	2049.4	2177.0	2283.5	2374.2	2452.7	2521.6
50000 S	1003.6	1421.6	1695.8	1896.5	2052.6	2179.6	2285.5	2375.5	2453.5	2522.0
50000 Fr	1003.7	1421.8	1695.6	1896.6	2053.3	2180.4	2286.4	2376.4	2454.3	2522.3
50000 Fn	1012.5	1439.9	1722.5	1931.5	2095.8	2229.8	2342.1	2438.2	2521.6	2594.9

	50000	100000	150000	200000	250000	300000	350000	400000	450000	500000
50000 S	2521.5	2922.0	3110.5	3221.1	3293.6	3345.3	3383.5	3413.0	3436.5	3455.4
50000 Fr	2522.3	2922.9	3110.8	3221.4	3294.5	3346.4	3385.0	3414.8	3438.3	3457.4
50000 Fn	2594.9	3034.1	3246.6	3374.9	3461.8	3525.1	3573.5	3611.9	3643.4	3669.9
100000 S	2922.9	3511.8	3810.0	3993.7	4119.1	4209.0	4277.1	4330.3	4372.9	4407.6
100000 Fr	2922.9	3511.4	3809.8	3993.5	4118.5	4209.0	4277.4	4330.8	4373.4	4408.1
100000 Fn	3034.1	3700.5	4054.9	4281.4	4440.8	4560.0	4653.2	4728.5	4790.9	4843.9
150000 S	3110.7	3810.6	4182.6	4417.0	4581.0	4701.5	4793.6	4865.8	4924.2	4971.6
150000 Fr	3110.8	3809.8	4181.3	4416.9	4580.5	4700.8	4792.8	4865.2	4923.5	4971.2
150000 Fn	3246.6	4054.9	4511.8	4816.2	5037.1	5206.3	5341.0	5451.5	5544.3	5623.9
200000 S	3222.8	3994.8	4418.5	4692.9	4886.1	5029.3	5140.4	5228.5	5299.7	5358.2
200000 Fr	3221.4	3993.5	4416.9	4691.1	4884.5	5028.4	5139.4	5227.5	5298.7	5357.4
200000 Fn	3374.9	4281.4	4816.2	5184.1	5458.0	5672.1	5845.6	5989.9	6112.6	6218.9
250000 S	3294.7	4118.7	4580.6	4885.2	5102.5	5264.5	5391.3	5492.9	5575.2	5642.7
250000 Fr	3294.5	4118.5	4580.5	4884.5	5101.5	5264.4	5391.1	5492.2	5574.4	5642.3
250000 Fn	3461.8	4440.8	5037.1	5458.0	5778.0	6032.8	6242.5	6419.3	6571.4	6704.5
300000 S	3346.1	4209.3	4701.3	5028.3	5264.6	5443.9	5583.6	5695.5	5787.5	5863.4
300000 Fr	3346.4	4209.0	4700.8	5028.4	5264.4	5443.0	5582.7	5694.7	5786.1	5862.0
300000 Fn	3525.1	4560.0	5206.3	5672.1	6032.8	6324.7	6568.3	6776.4	6957.5	7117.4
350000 S	3385.5	4277.9	4791.9	5139.3	5391.1	5582.6	5733.3	5854.5	5953.8	6036.3
350000 Fr	3385.0	4277.4	4792.8	5139.4	5391.1	5582.7	5733.3	5854.6	5954.0	6036.6
350000 Fn	3573.5	4653.2	5341.0	5845.6	6242.5	6568.3	6843.9	7082.1	7291.6	7478.4
400000 S	3416.0	4332.9	4866.7	5228.8	5493.5	5695.7	5855.7	5985.1	6091.4	6179.9
400000 Fr	3414.8	4330.8	4865.2	5227.5	5492.2	5694.7	5854.6	5983.8	6090.0	6178.6
400000 Fn	3611.9	4728.5	5451.5	5989.9	6419.3	6776.4	7082.1	7349.3	7586.8	7800.7
450000 S	3439.0	4373.1	4923.0	5297.7	5573.5	5785.6	5953.4	6089.9	6202.2	6295.8
450000 Fr	3438.3	4373.4	4923.5	5298.7	5574.4	5786.1	5954.0	6090.0	6202.1	6295.8
450000 Fn	3643.4	4790.9	5544.3	6112.6	6571.4	6957.5	7291.6	7586.8	7851.8	8092.8
500000 S	3457.4	4409.5	4972.9	5359.0	5643.0	5862.3	6036.8	6178.1	6296.2	6394.7
500000 Fr	3457.4	4408.1	4971.2	5357.4	5642.3	5862.0	6036.6	6178.6	6295.8	6394.0
500000 Fn	3669.9	4843.9	5623.9	6218.9	6704.5	7117.4	7478.4	7800.7	8092.8	8361.1

Tables 3.1 - 3.4: Difference of expected value (Fr, replacement model) and simulated mean (S) in percentage of the latter

	50	100	150	200	250	300	350	400	450	500
50	-1.5	-0.6	-0.2	0.1	0.2	-0.1	-0.2	-0.3	-0.4	-0.4
100	-1.2	-0.4	0.0	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	0.0
150	-0.8	-0.2	0.2	0.2	-0.2	-0.1	-0.2	-0.3	-0.4	-0.4
200	-0.3	-0.7	-0.2	0.0	-0.1	0.0	0.0	0.2	0.2	0.1
250	-1.2	-0.5	-0.2	0.1	0.0	-0.1	-0.1	-0.2	-0.2	-0.1
300	-0.3	-0.6	-0.2	0.1	-0.1	-0.2	0.0	0.1	0.0	0.1
350	-0.3	-0.1	-0.1	0.0	-0.1	-0.1	-0.1	-0.1	-0.1	0.0
400	-0.7	-0.2	0.1	0.2	0.0	0.0	0.0	0.1	0.2	0.2
450	0.4	0.4	0.5	0.2	0.0	0.1	0.1	0.1	0.1	0.1
500	-0.4	0.0	0.2	0.1	-0.1	-0.2	-0.1	-0.1	-0.2	-0.1

	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
500	-0.2	-0.1	-0.1	0.0	-0.1	0.0	-0.1	-0.1	0.0	0.0
1000	-0.2	-0.3	-0.2	-0.2	0.0	0.0	-0.1	0.0	0.0	-0.1
1500	-0.1	0.0	0.1	0.1	0.0	0.1	0.0	-0.1	-0.1	0.0
2000	-0.2	-0.1	-0.1	-0.2	-0.2	-0.1	-0.1	0.0	-0.1	-0.1
2500	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3000	-0.2	-0.1	-0.1	-0.1	-0.1	0.0	0.0	-0.1	-0.1	0.0
3500	0.1	0.0	0.0	-0.1	-0.1	0.0	0.0	0.0	0.0	-0.1
4000	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1
4500	-0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
5000	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.1	0.1

	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
5000	0.0	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2
10000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15000	-0.1	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	0.0
20000	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1
25000	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30000	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35000	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0
40000	-0.1	-0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
45000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

	50000	100000	150000	200000	250000	300000	350000	400000	450000	500000
50000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
150000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
200000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
250000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
300000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
350000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
400000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
450000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
500000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tables 3.5 - 3.8: Difference of expected value (Fn, non-replacement model) and simulated mean (S) in percentage of the latter

	50	100	150	200	250	300	350	400	450	500
50	-1.6	-0.6	-0.2	0.1	0.2	-0.1	-0.1	-0.3	-0.4	-0.4
100	-1.2	-0.4	0.0	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	0.0
150	-0.8	-0.2	0.2	0.2	-0.1	-0.1	-0.1	-0.3	-0.4	-0.4
200	-0.3	-0.7	-0.2	0.0	0.0	0.0	0.1	0.2	0.3	0.1
250	-1.1	-0.5	-0.2	0.1	0.1	0.0	-0.1	-0.1	-0.2	-0.1
300	-0.3	-0.6	-0.2	0.1	-0.1	-0.2	0.1	0.1	0.1	0.1
350	-0.3	-0.1	-0.1	0.0	-0.1	-0.1	0.0	-0.1	0.0	0.1
400	-0.7	-0.2	0.1	0.2	0.1	0.0	0.1	0.1	0.2	0.2
450	0.4	0.4	0.5	0.3	0.1	0.2	0.1	0.1	0.2	0.2
500	-0.3	0.0	0.2	0.1	-0.1	-0.1	0.0	0.0	-0.1	-0.1

	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
500	-0.1	-0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.1
1000	-0.2	-0.2	-0.1	-0.1	0.1	0.1	0.1	0.1	0.2	0.1
1500	0.0	0.1	0.2	0.2	0.2	0.3	0.2	0.1	0.1	0.2
2000	-0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.2
2500	0.3	0.2	0.3	0.2	0.2	0.3	0.3	0.3	0.3	0.3
3000	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.3
3500	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3
4000	0.4	0.3	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.3
4500	0.1	0.1	0.2	0.3	0.3	0.4	0.4	0.4	0.4	0.4
5000	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.5

	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
5000	0.4	0.6	0.7	0.8	0.9	1.0	1.0	1.0	1.0	1.0
10000	0.5	0.7	0.8	1.0	1.0	1.1	1.2	1.2	1.3	1.3
15000	0.5	0.8	1.0	1.1	1.2	1.3	1.3	1.4	1.5	1.6
20000	0.7	0.9	1.1	1.3	1.4	1.6	1.7	1.8	1.8	1.9
25000	0.7	1.1	1.3	1.5	1.6	1.7	1.8	1.9	2.0	2.1
30000	0.8	1.1	1.4	1.5	1.7	1.9	2.0	2.1	2.2	2.3
35000	0.9	1.3	1.5	1.7	1.9	2.1	2.2	2.3	2.4	2.5
40000	0.8	1.1	1.5	1.7	1.9	2.1	2.3	2.4	2.5	2.6
45000	0.9	1.2	1.5	1.7	2.0	2.2	2.3	2.5	2.6	2.7
50000	0.9	1.3	1.6	1.8	2.1	2.3	2.5	2.6	2.8	2.9

	50000	100000	150000	200000	250000	300000	350000	400000	450000	500000
50000	2.9	3.8	4.4	4.8	5.1	5.4	5.6	5.8	6.0	6.2
100000	3.8	5.4	6.4	7.2	7.8	8.3	8.8	9.2	9.6	9.9
150000	4.4	6.4	7.9	9.0	10.0	10.7	11.4	12.0	12.6	13.1
200000	4.7	7.2	9.0	10.5	11.7	12.8	13.7	14.6	15.3	16.1
250000	5.1	7.8	10.0	11.7	13.2	14.6	15.8	16.9	17.9	18.8
300000	5.4	8.3	10.7	12.8	14.6	16.2	17.6	19.0	20.2	21.4
350000	5.6	8.8	11.5	13.7	15.8	17.7	19.4	21.0	22.5	23.9
400000	5.7	9.1	12.0	14.6	16.9	19.0	20.9	22.8	24.5	26.2
450000	5.9	9.6	12.6	15.4	17.9	20.3	22.5	24.6	26.6	28.5
500000	6.1	9.9	13.1	16.0	18.8	21.4	23.9	26.3	28.5	30.8

Tables 4.1 - 4.2: Standard deviation for vocabulary overlap according to simulation (S) and according to formulae (Fr and Fn)

	50	100	150	200	250	300	350	400	450	500
50 S	2.0	2.4	2.6	2.7	2.8	2.9	3.0	3.0	3.1	3.1
50 Fr	2.2	2.6	2.9	3.1	3.2	3.4	3.5	3.6	3.6	3.7
50 Fn	2.2	2.6	2.9	3.1	3.2	3.4	3.5	3.6	3.6	3.7
100 S	2.4	2.8	3.1	3.2	3.4	3.5	3.5	3.6	3.6	3.7
100 Fr	2.6	3.1	3.4	3.7	3.8	4.0	4.1	4.3	4.4	4.5
100 Fn	2.6	3.1	3.4	3.7	3.8	4.0	4.1	4.3	4.4	4.5
150 S	2.3	2.9	3.3	3.5	3.7	3.8	3.9	4.1	4.2	4.2
150 Fr	2.9	3.4	3.8	4.0	4.2	4.4	4.5	4.7	4.8	4.9
150 Fn	2.9	3.4	3.8	4.0	4.2	4.4	4.5	4.7	4.8	4.9
200 S	2.6	3.3	3.6	3.9	4.0	4.2	4.4	4.4	4.5	4.6
200 Fr	3.1	3.7	4.0	4.3	4.5	4.7	4.8	5.0	5.1	5.2
200 Fn	3.1	3.7	4.0	4.3	4.5	4.7	4.8	5.0	5.1	5.2
250 S	2.8	3.5	3.7	3.9	4.1	4.3	4.4	4.6	4.7	4.7
250 Fr	3.2	3.8	4.2	4.5	4.7	4.9	5.0	5.2	5.3	5.4
250 Fn	3.2	3.8	4.2	4.5	4.7	4.9	5.0	5.2	5.3	5.4
300 S	2.7	3.6	4.0	4.4	4.7	5.0	5.0	5.1	5.2	5.2
300 Fr	3.4	4.0	4.4	4.7	4.9	5.1	5.2	5.4	5.5	5.6
300 Fn	3.4	4.0	4.4	4.7	4.9	5.1	5.2	5.4	5.5	5.6
350 S	2.8	3.5	4.0	4.2	4.5	4.7	5.0	5.1	5.3	5.4
350 Fr	3.5	4.1	4.5	4.8	5.0	5.2	5.4	5.6	5.7	5.8
350 Fn	3.5	4.1	4.5	4.8	5.0	5.2	5.4	5.6	5.7	5.8
400 S	2.8	3.6	4.0	4.4	4.8	4.9	5.1	5.3	5.5	5.6
400 Fr	3.6	4.3	4.7	5.0	5.2	5.4	5.6	5.7	5.9	6.0
400 Fn	3.6	4.3	4.7	5.0	5.2	5.4	5.6	5.7	5.9	6.0
450 S	3.0	3.9	4.3	4.6	4.8	5.1	5.2	5.3	5.4	5.6
450 Fr	3.6	4.4	4.8	5.1	5.3	5.5	5.7	5.9	6.0	6.1
450 Fn	3.6	4.4	4.8	5.1	5.3	5.5	5.7	5.9	6.0	6.1
500 S	3.0	3.8	4.3	4.6	4.9	5.2	5.5	5.8	5.8	6.1
500 Fr	3.7	4.5	4.9	5.2	5.4	5.6	5.8	6.0	6.1	6.3
500 Fn	3.7	4.5	4.9	5.2	5.4	5.6	5.8	6.0	6.1	6.3

	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
500 S	5.8	6.6	6.8	7.1	7.2	7.4	7.5	7.7	7.9	7.9
500 Fr	6.3	7.3	7.9	8.4	8.8	9.0	9.3	9.5	9.7	9.8
500 Fn	6.3	7.3	7.9	8.4	8.8	9.1	9.3	9.5	9.7	9.8
1000 S	6.5	7.6	8.4	8.7	8.9	9.4	9.6	9.7	9.8	9.8
1000 Fr	7.3	8.3	9.1	9.6	10.1	10.4	10.7	11.0	11.3	11.5
1000 Fn	7.3	8.4	9.1	9.6	10.1	10.4	10.8	11.0	11.3	11.5
1500 S	6.8	8.1	8.9	9.5	9.9	10.2	10.4	10.8	11.2	11.5
1500 Fr	7.9	9.1	9.8	10.4	10.8	11.2	11.6	11.9	12.2	12.4
1500 Fn	7.9	9.1	9.8	10.4	10.9	11.3	11.6	11.9	12.2	12.4
2000 S	7.1	8.5	9.3	10.1	10.5	11.0	11.2	11.6	11.9	12.1
2000 Fr	8.4	9.6	10.4	11.0	11.4	11.8	12.2	12.5	12.8	13.0
2000 Fn	8.4	9.6	10.4	11.0	11.4	11.8	12.2	12.5	12.8	13.0
2500 S	7.8	9.1	10.1	10.7	11.0	11.4	11.9	12.3	12.4	12.6
2500 Fr	8.8	10.1	10.8	11.4	11.9	12.3	12.7	13.0	13.2	13.5
2500 Fn	8.8	10.1	10.9	11.4	11.9	12.3	12.7	13.0	13.3	13.5
3000 S	7.7	9.2	10.2	10.7	11.3	11.8	12.0	12.3	12.7	13.0
3000 Fr	9.0	10.4	11.2	11.8	12.3	12.7	13.1	13.4	13.6	13.9
3000 Fn	9.1	10.4	11.3	11.8	12.3	12.7	13.1	13.4	13.7	13.9
3500 S	8.2	9.7	10.8	11.7	12.1	12.4	12.8	13.2	13.4	13.8
3500 Fr	9.3	10.7	11.6	12.2	12.7	13.1	13.4	13.7	14.0	14.2
3500 Fn	9.3	10.8	11.6	12.2	12.7	13.1	13.4	13.7	14.0	14.2
4000 S	8.3	9.9	10.8	11.7	12.4	12.9	13.3	13.6	13.9	14.3
4000 Fr	9.5	11.0	11.9	12.5	13.0	13.4	13.7	14.0	14.3	14.5
4000 Fn	9.5	11.0	11.9	12.5	13.0	13.4	13.7	14.0	14.3	14.5
4500 S	8.3	10.2	11.4	12.2	12.6	13.0	13.2	13.7	14.1	14.3
4500 Fr	9.7	11.3	12.2	12.8	13.2	13.6	14.0	14.3	14.5	14.8
4500 Fn	9.7	11.3	12.2	12.8	13.3	13.7	14.0	14.3	14.6	14.8
5000 S	8.3	10.3	11.9	12.7	12.9	13.3	13.5	13.7	14.0	14.0
5000 Fr	9.8	11.5	12.4	13.0	13.5	13.9	14.2	14.5	14.8	15.0
5000 Fn	9.8	11.5	12.4	13.0	13.5	13.9	14.2	14.5	14.8	15.0

Tables 4.3 - 4.4: Standard deviation for vocabulary overlap according to simulation (S) and according to formulae (Fr and Fn)

	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
5000 S	14.3	15.4	15.8	16.5	17.1	17.3	17.7	18.0	18.0	18.2
5000 Fr	15.0	16.7	17.7	18.5	19.0	19.5	19.8	20.1	20.4	20.6
5000 Fn	15.0	16.7	17.8	18.5	19.1	19.6	19.9	20.3	20.5	20.7
10000 S	15.6	17.0	18.4	18.9	19.4	20.2	20.5	20.6	21.0	21.1
10000 Fr	16.7	18.1	19.0	19.7	20.3	20.9	21.3	21.7	22.0	22.3
10000 Fn	16.7	18.1	19.0	19.8	20.4	21.0	21.4	21.8	22.2	22.5
15000 S	16.9	18.1	19.2	19.4	19.6	20.4	20.6	20.7	21.2	21.7
15000 Fr	17.7	19.0	19.8	20.4	21.0	21.5	21.9	22.3	22.6	23.0
15000 Fn	17.8	19.0	19.8	20.5	21.1	21.6	22.0	22.5	22.8	23.2
20000 S	17.1	19.1	19.6	20.3	20.8	21.2	21.7	21.8	22.2	22.8
20000 Fr	18.5	19.7	20.4	21.0	21.5	21.9	22.3	22.7	23.0	23.4
20000 Fn	18.5	19.8	20.5	21.0	21.6	22.0	22.5	22.9	23.2	23.6
25000 S	17.4	19.7	20.5	20.4	21.3	22.2	22.3	22.4	23.1	23.3
25000 Fr	19.0	20.3	21.0	21.5	21.9	22.3	22.7	23.0	23.3	23.7
25000 Fn	19.1	20.4	21.1	21.6	22.0	22.4	22.8	23.2	23.5	23.9
30000 S	17.9	19.9	20.6	21.1	21.8	21.9	22.3	23.0	23.9	24.4
30000 Fr	19.5	20.9	21.5	21.9	22.3	22.7	23.0	23.3	23.6	23.9
30000 Fn	19.6	21.0	21.6	22.0	22.4	22.8	23.1	23.5	23.8	24.1
35000 S	18.2	20.2	20.4	21.9	22.2	22.8	23.1	23.9	23.8	24.1
35000 Fr	19.8	21.3	21.9	22.3	22.7	23.0	23.3	23.6	23.9	24.1
35000 Fn	19.9	21.4	22.0	22.5	22.8	23.1	23.5	23.8	24.1	24.3
40000 S	17.9	21.2	21.8	22.3	22.8	22.9	23.2	23.4	23.6	23.9
40000 Fr	20.1	21.7	22.3	22.7	23.0	23.3	23.6	23.9	24.1	24.3
40000 Fn	20.3	21.8	22.5	22.9	23.2	23.5	23.8	24.0	24.3	24.6
45000 S	18.7	21.1	21.6	22.4	22.6	22.9	23.0	23.7	24.1	24.5
45000 Fr	20.4	22.0	22.6	23.0	23.3	23.6	23.9	24.1	24.3	24.6
45000 Fn	20.5	22.2	22.8	23.2	23.5	23.8	24.1	24.3	24.5	24.8
50000 S	19.0	21.1	22.5	23.2	23.3	23.2	23.4	23.3	23.9	24.7
50000 Fr	20.6	22.3	23.0	23.4	23.7	23.9	24.1	24.3	24.6	24.8
50000 Fn	20.7	22.5	23.2	23.6	23.9	24.1	24.3	24.6	24.8	25.0

	50000	100000	150000	200000	250000	300000	350000	400000	450000	500000
50000 S	24.1	26.3	26.9	27.7	28.3	28.9	29.2	29.6	30.2	30.4
50000 Fr	24.8	26.5	27.8	28.7	29.3	29.8	30.2	30.6	30.8	31.0
50000 Fn	25.0	26.9	28.4	29.5	30.4	31.0	31.6	32.0	32.4	32.7
100000 S	25.3	26.9	27.9	28.9	29.9	30.1	30.7	30.6	30.8	31.0
100000 Fr	26.5	27.4	28.2	28.9	29.5	30.1	30.5	30.9	31.2	31.5
100000 Fn	26.9	27.7	28.7	29.6	30.5	31.3	32.0	32.6	33.1	33.6
150000 S	27.7	27.1	29.0	29.6	30.2	30.7	31.2	31.2	31.6	31.7
150000 Fr	27.8	28.2	28.6	29.0	29.5	29.9	30.2	30.6	30.9	31.1
150000 Fn	28.4	28.7	29.0	29.5	30.1	30.7	31.3	31.9	32.4	32.8
200000 S	28.1	28.2	28.4	28.4	28.3	29.1	29.9	30.0	30.3	30.7
200000 Fr	28.7	28.9	29.0	29.2	29.4	29.7	30.0	30.2	30.4	30.6
200000 Fn	29.5	29.6	29.5	29.5	29.7	30.0	30.3	30.7	31.0	31.2
250000 S	28.8	29.3	29.0	29.7	29.0	28.1	28.5	29.0	29.6	29.7
250000 Fr	29.3	29.5	29.5	29.4	29.5	29.6	29.7	29.9	30.0	30.1
250000 Fn	30.4	30.5	30.1	29.7	29.5	29.3	29.2	29.1	29.0	28.8
300000 S	30.6	29.4	29.3	29.6	29.3	29.5	29.4	29.3	30.0	30.0
300000 Fr	29.8	30.1	29.9	29.7	29.6	29.6	29.6	29.6	29.7	29.7
300000 Fn	31.0	31.3	30.7	30.0	29.3	28.6	27.9	27.2	26.5	25.5
350000 S	29.1	30.2	30.9	31.8	31.4	31.6	30.8	30.5	30.4	29.7
350000 Fr	30.2	30.5	30.2	30.0	29.7	29.6	29.5	29.4	29.4	29.3
350000 Fn	31.6	32.0	31.3	30.3	29.2	27.9	26.6	25.0	23.2	21.0
400000 S	29.8	30.8	30.3	30.6	30.1	29.6	29.6	29.2	29.2	29.3
400000 Fr	30.6	30.9	30.6	30.2	29.9	29.6	29.4	29.2	29.1	29.0
400000 Fn	32.0	32.6	31.9	30.7	29.1	27.2	25.0	22.3	18.9	14.0
450000 S	29.9	31.7	30.3	30.1	29.2	29.4	29.1	29.1	28.9	28.7
450000 Fr	30.8	31.2	30.9	30.4	30.0	29.7	29.4	29.1	28.9	28.7
450000 Fn	32.4	33.1	32.4	31.0	29.0	26.5	23.2	18.9	12.2	nan
500000 S	31.3	30.9	30.6	30.3	29.5	28.7	28.3	27.9	27.9	27.9
500000 Fr	31.0	31.5	31.1	30.6	30.1	29.7	29.3	29.0	28.7	28.4
500000 Fn	32.7	33.6	32.8	31.2	28.8	25.5	21.0	14.0	nan	nan

Tables 5.1 - 5.4: Difference of theoretical (Fr) and ps.-empirical (S) standard deviation in percentage of the latter

	50	100	150	200	250	300	350	400	450	500
50	8.3	10.5	12.2	15.4	16.6	16.2	16.4	16.8	17.9	20.0
100	8.9	10.4	12.3	13.3	12.7	15.9	18.0	18.5	20.1	20.8
150	23.8	18.9	13.6	14.5	14.0	14.0	15.3	14.0	14.2	16.2
200	18.4	11.7	10.3	10.4	10.5	10.1	9.5	12.2	12.7	14.0
250	14.6	9.4	14.7	13.8	13.9	13.6	13.5	13.8	13.7	14.8
300	23.0	12.4	8.3	5.5	4.7	1.2	3.9	5.8	6.0	8.0
350	21.6	16.9	13.9	13.4	12.2	11.0	8.2	8.1	6.7	7.2
400	25.4	18.2	17.8	12.4	9.0	9.7	9.3	7.5	6.7	6.2
450	19.1	13.1	10.8	10.3	10.0	8.7	10.5	11.4	11.8	9.7
500	23.0	17.8	12.7	13.8	10.8	9.0	5.4	3.7	4.9	3.0

	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
500	7.8	10.1	15.8	18.7	22.1	22.9	23.9	23.2	23.3	24.1
1000	12.3	10.1	8.5	10.2	12.6	11.1	12.2	13.2	14.8	16.7
1500	16.9	12.2	10.0	9.1	9.3	9.9	11.2	10.2	8.9	7.6
2000	17.8	13.1	11.4	8.3	8.6	7.8	8.5	7.6	7.6	7.8
2500	11.8	11.0	7.5	7.2	8.6	7.8	6.0	5.6	6.7	7.5
3000	18.2	13.0	10.2	10.3	9.0	7.9	8.5	8.8	7.5	6.9
3500	12.8	10.7	7.3	3.7	4.4	5.0	4.4	4.2	4.0	2.8
4000	14.9	11.5	10.5	7.0	4.3	4.0	2.9	3.0	2.8	1.4
4500	16.8	10.9	6.4	4.7	5.1	4.9	5.6	3.9	3.3	2.9
5000	17.9	11.2	4.2	2.3	4.6	4.4	5.5	5.8	5.7	7.3

	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
5000	4.6	8.2	12.1	12.1	11.6	12.6	12.0	12.0	13.3	13.3
10000	6.6	6.2	3.4	4.1	4.6	3.3	3.8	5.3	4.9	5.6
15000	5.1	5.0	2.8	5.1	7.0	5.0	6.5	7.8	6.9	5.8
20000	8.3	3.5	4.0	3.1	3.2	3.3	3.0	3.9	4.0	2.3
25000	9.5	3.4	2.4	5.1	2.6	0.6	1.8	2.8	0.9	1.6
30000	8.9	4.6	4.3	3.9	2.4	3.3	2.9	1.2	-1.1	-2.1
35000	8.8	5.5	7.2	1.8	2.1	0.7	0.9	-1.4	0.3	0.2
40000	12.2	2.2	2.4	1.6	0.9	1.8	1.5	1.7	2.2	1.9
45000	8.8	4.2	4.9	2.7	3.1	3.1	3.9	1.6	0.9	0.2
50000	8.2	5.8	2.2	0.8	1.6	2.8	3.2	4.4	2.8	0.4

	50000	100000	150000	200000	250000	300000	350000	400000	450000	500000
50000	2.6	0.8	3.4	3.6	3.6	3.4	3.7	3.2	2.1	2.2
100000	4.7	1.6	1.0	0.1	-1.1	-0.1	-0.7	0.9	1.3	1.3
150000	0.2	3.9	-1.4	-2.0	-2.4	-2.8	-3.1	-1.9	-2.5	-2.0
200000	2.1	2.7	2.1	2.9	4.1	2.2	0.0	0.5	0.5	-0.3
250000	1.8	0.9	1.5	-0.9	1.7	5.4	4.4	3.0	1.4	1.5
300000	-2.5	2.2	1.9	0.5	1.1	0.2	0.7	1.0	-1.2	-1.0
350000	4.0	0.9	-2.0	-5.7	-5.4	-6.3	-4.2	-3.5	-3.3	-1.3
400000	2.6	0.1	0.9	-1.4	-0.7	0.0	-0.8	0.1	-0.3	-1.2
450000	3.0	-1.5	1.8	1.2	2.7	1.1	0.8	0.1	-0.2	0.1
500000	-0.8	1.8	1.5	1.1	2.3	3.4	3.8	4.0	2.7	1.8

Tables 5.5 - 5.8: Difference of theoretical (Fn) and ps.-empirical (S) standard deviation in percentage of the latter

	50	100	150	200	250	300	350	400	450	500
50	8.3	10.5	12.2	15.4	16.6	16.2	16.4	16.8	17.9	20.0
100	8.9	10.4	12.3	13.3	12.7	15.9	18.0	18.5	20.1	20.8
150	23.8	18.9	13.6	14.6	14.1	14.0	15.3	14.0	14.2	16.2
200	18.4	11.7	10.3	10.4	10.6	10.1	9.6	12.2	12.7	14.0
250	14.6	9.4	14.7	13.8	13.9	13.6	13.5	13.9	13.7	14.8
300	23.1	12.4	8.3	5.5	4.7	1.2	3.9	5.8	6.1	8.1
350	21.6	16.9	13.9	13.4	12.2	11.0	8.2	8.1	6.7	7.2
400	25.4	18.2	17.8	12.4	9.0	9.7	9.4	7.5	6.8	6.2
450	19.1	13.1	10.8	10.3	10.0	8.7	10.5	11.4	11.8	9.7
500	23.0	17.8	12.7	13.8	10.9	9.0	5.4	3.7	4.9	3.1

	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
500	7.8	10.2	15.8	18.7	22.2	23.0	24.0	23.3	23.4	24.2
1000	12.3	10.2	8.5	10.2	12.7	11.2	12.3	13.4	14.9	16.8
1500	16.9	12.2	10.0	9.2	9.3	10.0	11.3	10.4	9.0	7.7
2000	17.8	13.2	11.5	8.4	8.7	8.0	8.6	7.7	7.7	8.0
2500	11.8	11.1	7.6	7.3	8.7	7.9	6.1	5.8	6.8	7.7
3000	18.3	13.1	10.3	10.4	9.1	8.0	8.6	8.9	7.7	7.0
3500	12.9	10.8	7.4	3.9	4.5	5.1	4.5	4.3	4.1	2.9
4000	15.0	11.6	10.6	7.1	4.4	4.1	3.1	3.2	3.0	1.6
4500	16.9	11.0	6.6	4.8	5.3	5.1	5.8	4.0	3.5	3.1
5000	18.0	11.3	4.3	2.5	4.7	4.6	5.7	5.9	5.9	7.5

	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
5000	4.8	8.5	12.4	12.5	12.0	13.2	12.5	12.6	13.9	14.0
10000	6.8	6.5	3.7	4.5	5.1	3.8	4.4	6.0	5.7	6.5
15000	5.4	5.3	3.1	5.5	7.5	5.6	7.1	8.5	7.8	6.7
20000	8.7	3.9	4.5	3.6	3.7	3.9	3.7	4.6	4.8	3.1
25000	10.0	3.9	2.9	5.6	3.1	1.2	2.4	3.5	1.7	2.5
30000	9.4	5.2	4.8	4.5	3.0	3.9	3.6	1.9	-0.3	-1.2
35000	9.4	6.2	7.8	2.5	2.7	1.4	1.6	-0.6	1.1	1.0
40000	12.9	2.9	3.1	2.3	1.6	2.5	2.2	2.5	3.1	2.7
45000	9.4	4.9	5.7	3.5	3.9	4.0	4.7	2.4	1.7	1.1
50000	8.8	6.7	3.1	1.6	2.5	3.7	4.0	5.3	3.7	1.4

	50000	100000	150000	200000	250000	300000	350000	400000	450000	500000
50000	3.5	2.3	5.7	6.7	7.2	7.6	8.3	8.2	7.4	7.8
100000	6.3	3.0	2.8	2.6	2.2	4.0	4.2	6.6	7.6	8.2
150000	2.4	5.8	0.1	-0.4	-0.3	0.0	0.4	2.4	2.4	3.4
200000	5.1	5.3	3.8	4.1	5.1	3.3	1.3	2.1	2.3	1.5
250000	5.4	4.2	3.8	0.0	1.5	4.2	2.4	0.3	-2.0	-3.1
300000	1.4	6.4	4.9	1.5	-0.1	-3.2	-4.9	-7.2	-11.9	-15.0
350000	8.6	5.9	1.5	-4.5	-7.2	-11.5	-13.7	-17.9	-23.6	-29.4
400000	7.6	5.7	5.3	0.1	-3.3	-8.0	-15.6	-23.6	-35.4	-52.4
450000	8.3	4.6	6.9	3.0	-0.8	-9.9	-20.4	-35.1	-57.8	nan
500000	4.7	8.7	7.1	3.0	-2.3	-11.2	-25.8	-49.9	nan	nan

Table 6:
Mean vocabulary overlap, based on corpora of lengths 50,000 to 450,000

	Length N' and vocabulary size V' of basic corpus									N && V
N'	50000	100000	150000	200000	250000	300000	350000	400000	450000	465744
V'	657.8	4814.2	5556.0	6118.8	6599.0	6993.0	7358.0	7666.4	7932.8	8011.0
m && n										
50000	2558.5	2574.4	2554.7	2543.5	2538.5	2534.1	2529.3	2525.2	2522.9	2521.5
100000	3250.6	3498.6	3522.5	3527.0	3526.9	3522.0	3517.8	3512.4	3511.8	3511.8
150000	3507.3	4021.5	4128.7	4165.7	4181.1	4184.5	4185.2	4181.7	4182.6	4182.6
200000	3601.9	4331.5	4538.3	4619.6	4660.3	4675.3	4684.7	4684.7	4690.0	4692.9
250000	3637.3	4521.0	4828.2	4960.1	5031.4	5063.3	5085.6	5091.4	5100.2	5102.5

m && n, in the leftmost column, are extract lengths (here equal).