Quantitative parameters in corpus design: Estimating the optimum text size in Modern Greek language

George Mikros

University of Athens, Department of Italian & Spanish Language and Literature & Institute for Language and Speech Processing
Panepistimioupoli Zografou 15784 Athens, Greece
gmikros@isll.uoa.gr

Abstract

The aim of this paper is to investigate the major quantitative parameters related to the definition of the optimum text size in Modern Greek corpus development. Using the Hellenic National Corpus (HNC) (Hatzigeorgiu et al., 2000) as a reference point we estimated a number of critical statistical measures regarding feature counting in different text sizes. The results indicate that frequent linguistic features behave differently from the medium frequency and the rare ones and the text size increase do not affect them uniformly.

1. Introduction

Most corpora, regarding the text size issue, use as a reference point the choices that were made in the design of the Brown corpus (Kucera & Frances, 1967). However it has been pointed out that significant choices in the Brown corpus design were based on the technological and socio-scientific circumstances of this era (Eagles, 1996: 3). Especially the issue of the text size (fixed in 2000+ words per text in Brown corpus) the last decade has begun to be questioned and a number of different text sizes has been adopted in recent corpora.

Although there are few empirical investigations regarding the issue of optimal sample size there is a general agreement that the occurrence of linguistic features is highly dependent from their frequency, as it is counted in a general language corpus, and the text size that is used to locate them. Previous studies (de Haan, 1992) have shown that frequent linguistic features do not correlate significantly with the text size and present stability and uniformity of occurrence across different text sizes. On the other hand, low frequency features exhibit considerable variation and need text sizes of considerable amount in order to stabilize their occurrence. Biber & Finegan (1991) examining data in English language analysed the distribution of a number of linguistic features across text sizes of 1000 words extracted from larger texts from the LOB and London - Lund corpora. They found that counts of frequent features are relative stable across 1000 words sample.

2. Methodology

2.1. Selection of the linguistic features

In order to define the optimum text size in a corpus of Modern Greek we used the HNC as a starting point. In the HNC there are texts of different sizes regarding mainly from the medium of the text. In order to maintain uniformity of medium and at the same time a possibility to use texts that vary significantly in size we used the newspaper portion of the corpus.

The effect of the text size to the corpus design was investigated adopting a user-oriented view. The search of specific linguistic units becomes the main activity of most corpora users. The main virtue of every corpus that is used in this way is the ability to respond equally well to queries of specific features that are rare with those that are frequent. Considering this as the main criterion for validating the results of our study, we tried to test the effect of different text sizes in the presence of each linguistic feature taking into consideration its overall frequency in the general language.

In order to have a representative sample of linguistic features, which would play the role of search words in the corpus, we selected 36 linguistic units, which were further subdivided according their linguistic function (phonological, morphological, and lexical). For each feature, we included its absolute frequency of occurrence in the general Modern Greek language corpus (HNC). In order to homogenize the frequency scale we transformed the raw frequency data of each feature into a discrete variable that defined the frequency of a linguistic feature in a scale from 1 to 9 (1 was the most frequent feature while 9 was the most rare).

The selection of the lexical features was based on a frequency word list of the whole HNC. This list was divided in nine frequency areas. From each area we randomly extracted two words of different lexical category (part of speech). Following the same random procedure we extracted nine phonological and morphological features. The features selection is based on the following assumptions:

a) Uniform distribution: In order to maintain a uniform distribution of the selected features we ensure that all the features of the same frequency area were closely matched regarding their absolute frequency of occurrence in the HNC.

b) Non-correlation with the topic and genre of the text: Before selecting a feature we examined the correlation matrixes constructed between the feature and the topic or genre of the texts correspondingly

The linguistic features that correspond to the above criteria are displayed in the table 1:

Frequency Area	Word	Frequency of occurrence in HNC	Phoneme	Frequency of occurrence in HNC	Morpheme	Frequency of occurrence in HNC
1	tou na	14684 13787	the	13487	-ous	11885
2	gia einai	7757 7371	sm	7069	-is	6672
3	tis ena	3413 2537	kl	3288	-os	4945
4	itan mou	1856 1839	sth	1849	-oume	2861
5	prepei xronia	1054 906	xth	904	-ete	1071
6	iparxoun politiki	472 433	lt	466	-eos	421
7	kivernisi apotelei	224 213	nst	196	-menis	187
8	dosei thalassa	100 100	dl	101	-simo	136
9	simiothei skepsi	50 50	sn	51	-ousas	50

Table 1: Frequency classification of the selected linguistic features

2.2. Construction of the test corpus

In order to study the effects of text size we constructed sample frames of different text sizes. We preserved a general categorization scheme, which categorizes texts in categories that were different approximately 500 words from each other (6 size categories in total).

The distribution of the texts based on this categorization is shown in table 2:

Text-size frame (in words)	Number of texts	Size category
100-500	86	1
500-1000	95	2
1000-1500	154	3
1500-2000	67	4
2000-3000	58	5
3000<	25	6

Table2: Text distribution among text-size frames

The total corpus consists of 651.000 words in 485 texts that cover a wide spectrum of text genres and text topics

The above test corpus was used in order to investigate the following research questions:

a) How many texts of the total corpus do not contain the linguistic feature that we are looking for and how this number correlate with the text size and the frequency of the feature?

b) How two basic statistic measures of distribution homogeneity (the standard deviation and the standard error of the mean) correlate to the text size that is used and the frequency of the feature that is searched?

The investigation of the above research questions does not provide direct answers or rules of thumb in the practice of corpus design since the concept of optimum text size cannot be defined without reference to a specific design framework. However, we can define some basic quantitative trends and exploit them in the text sampling procedure. These trends can help us especially in the newspaper texts which are characterized from wide text size variation. In that case we can locate the intercept between two opposite requirements in text collection:

a) Large texts which are rarer in newspapers, but embed a significant amount of different linguistic features

b) Small texts which are more frequent in newspapers but do not contain rare linguistic features.

3. Results

3.1. Percentage of non-appearance of the selected linguistics features (PNA)

The selected linguistic features were searched in the test corpus and their frequency of occurrence was recorded for each text file. For each one we calculated the percentage of non-appearance inside each text-size frame (as percentage of text files that do not exhibit a certain linguistic feature in relation to the number of total text files that exist in a specific text-size frame). This measure gave us an indication of the extent of the feature's presence in the test corpus.

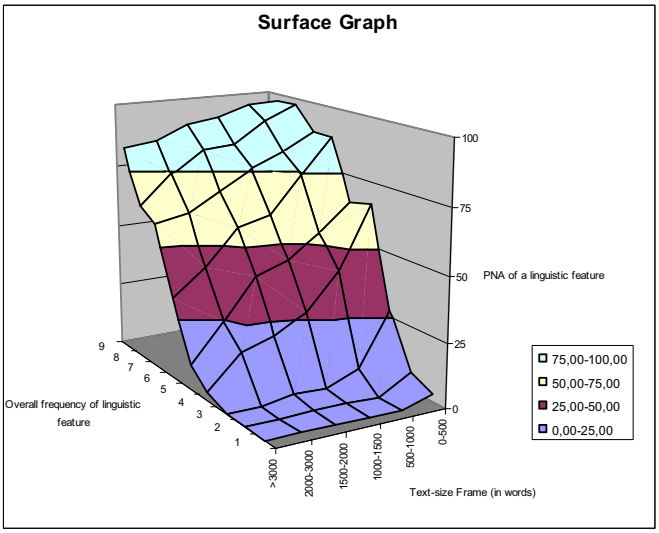
Investigating the percentage of non-appearance (PNA) of the selected features in relation to their overall frequency we conclude as was expected, that there is a perfect correlation between them. Enhancing our analysis we added the variable of the text size and we investigated the PNA of the features in a three-way table which combines the three main variables of our analysis: PNA of features in relation to their overall frequency and their text size frame in which they were

searched. A multiple regression analysis with the above variables exhibits a very strong R^2 (0,87). The results are displayed in the table 3:

Variables	*Stand. Beta*	*t*	*Sig.*
(Constant)		1,562	0,120
Frequency of Linguistic Feature	0,888	36,225	0,000
Size of Text	-0,285	-11,650	0,000
Linguistic Level of the Feature (phon, morph., lex.)	-0,055	-2,264	0,025

Table 3: Multiple regression results. Dependent variable: PNA

From the above analysis it is evident that the frequency of a linguistic feature is the most influencing factor regarding its appearance in a corpus. The second most influential factor appears to be the size of the text files that we included in our corpus. A third factor that reaches statistical significance is the linguistic level of the searched feature. However, its influence is restricted compared to the effect of the previous mentioned variables. The following graph (Graph 1) displays the influence of the independent variables of the above regression analysis:



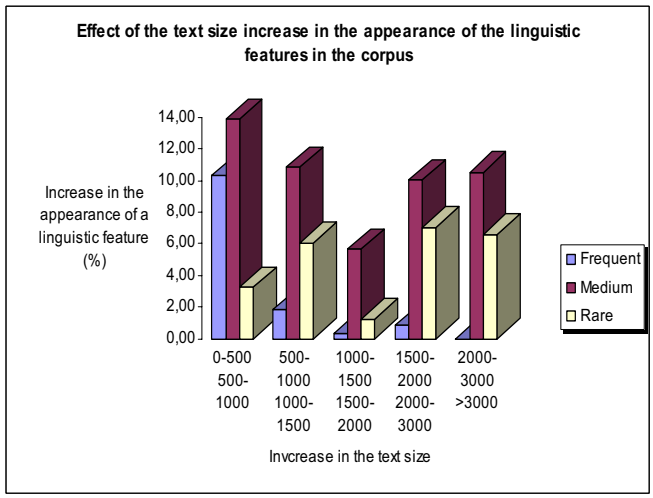
Graph 1: Interaction of text-size and frequency of a linguistic feature in its PNA

By examining the above graph we can reach certain conclusions regarding the interaction of the independent variables of our study:

- The frequent linguistic features (from the 1st up to the 3rd frequency area) exhibit stability in text sizes over 500 words. There is uniformity of appearance and their PNA does not correlate with the text size.
- The linguistic features that belong to the medium frequency area (from the 4th to the 6th frequency area) exhibit strong correlation with the text size. The probability to search for them and not to find them in the corpus is increasing linearly in relation to the size of the text we are performing the search.
- The rare linguistic features (from the 7th to the 9th frequency area) exhibit similar behavior with the frequent features. They do not correlate with the text size, since they cannot be found even if we are searching in very large text (> 3000 words).

The correlational analysis shows that the linguistic features of the medium frequency area exhibit considerable sensitivity in the variation of the text size. In order to examine further the effect of the text size in the appearance of searched items in a corpus, we need to define which increase in the text size has the most decisive influence on the PNA of a linguistic feature.

We investigated the above issue by comparing the PNA of the features sequentially. We took the PNA of the first text size frame in all the features and we compared it with the PNA of the next immediate text size category. Each text size frame differs from each other approximately 500 words. Consequently, we estimate the effect of 500 words increase in the PNA of the selected linguistic features. This effect is displayed in the following graph (Graph 2):



Graph 2: Augmentation of the appearance of a linguistic feature conditioned by the 500 word increment in text size

The above graph shows that the greatest improvement in the representation of frequent and medium frequency features is met when the text size is incrementing from the 500 to 1000 words. On the other hand, the lowest improvement in the appearance of a feature happens when we increment the text size from 1000 -1500 to 1500 – 2000. This is an indication that text sizes from 1000 to 2000 words exhibit stability and uniformity in the quantitative representation of the linguistic features regardless their overall frequency of occurrence.

Rare features behave differently. The highest rates of improvement are exhibited in the transition from the text sizes of 1500 – 2000 to 2000 – 3000 words. However, significant improvement is also noted in the transition from the 500 – 1000 to 1000 – 1500 text size frame.

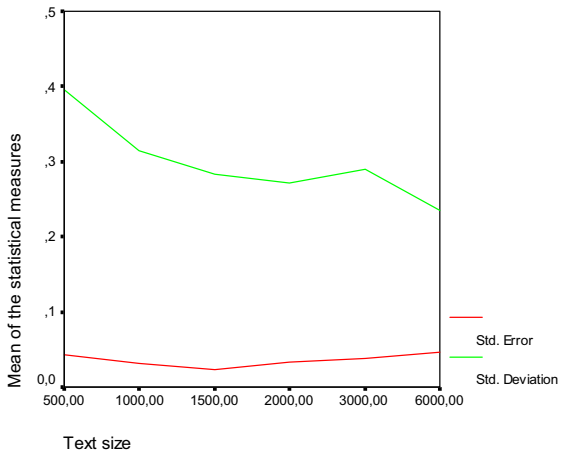
3.2. The influence of the text size in the standard deviation and standard error of the mean of the linguistic features

A different way to estimate the influence of the text size in corpus design is to keep track of the evolution of two well known statistical measures of data dispersion:

a) Standard deviation (SD): A measure of the range of variation between the mean occurrences of a linguistic feature.

b) Standard error of the mean (SEM): A measure of how much the value of the mean occurrence of a linguistic feature may vary from sample to sample taken from the same distribution.

Both measures were calculated for each linguistic feature in each text size frame separately. The results are displayed in graph 3:



Graph 3: Evolution of the SD and SEM in relation to the text size

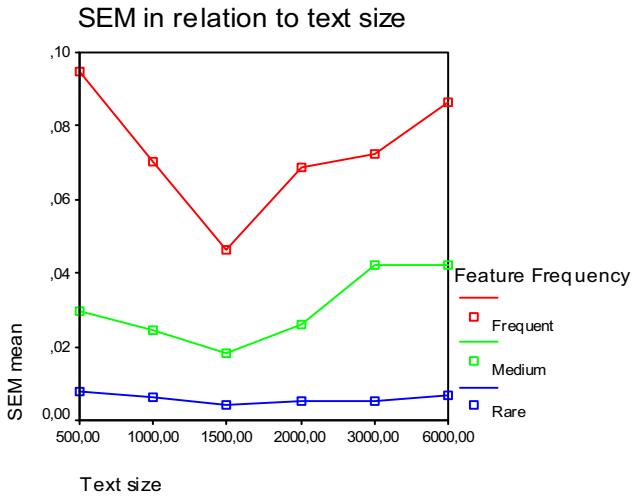
The data displayed in the graph support the view that both measures are negatively correlated with the text size. In order to have a more detailed view of the nature of the interaction of both statistical measures and the independent variables of our study we conducted two multiple regressions with dependent variable the SD and SEM correspondingly and independent variables the text size and the linguistic level of the feature. The results are displayed in Table 4:

Dependent variable: Standard Deviation			
Variables	***Stand. Beta***	***t***	***Sig.***
(Constant)		21,7	0,000
Frequency of Linguistic Feature	-0,831	-22,428	0,000
Size of Text	-0,125	-3,371	0,001
Linguistic Level of the Feature (phon, morph., lex.)	0,043	1,150	0,252
Dependent variable: Standard Error of the Mean			
Variables	***Stand. Beta***	***t***	***Sig.***
(Constant)		17,505	0,000
Frequency of Linguistic Feature	-0,803	-20,016	0,000
Size of Text	0,112	2,803	0,006
Linguistic Level of the Feature (phon, morph., lex.)	0,022	0,538	0,591

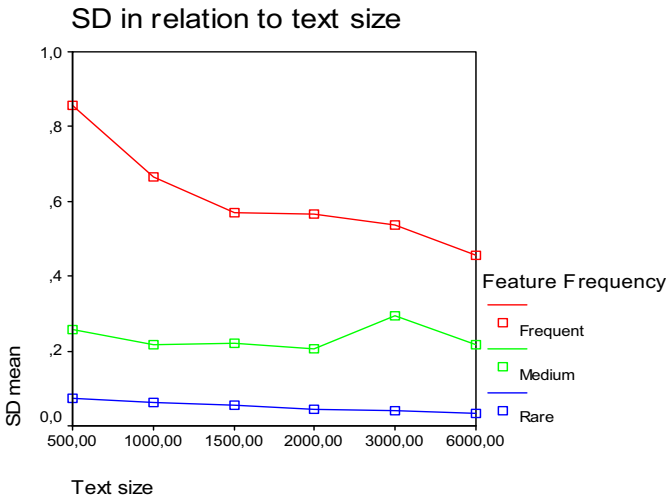
Table 4: Multiple regression results with dependent variables the SD and SEM

Both analyses confirm that the most influencing factor on the SD and SEM of the features is the frequency of the linguistic feature itself. Second most influential factor is the text size while the linguistic level of the feature does not reach statistical significance.

The interaction of the above significant factors results in a conditioned variation of the SEM and SD of the linguistic features and is displayed in the following graphs (Graph 4 & 5):



Graph 4: Evolution of the SEM in relation to the text size and the frequency of the linguistic feature

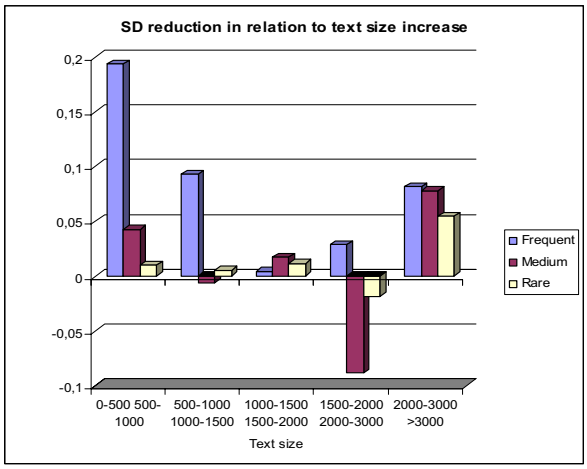


Graph 5: Evolution of the SD in relation to the text size and the frequency of the linguistic feature

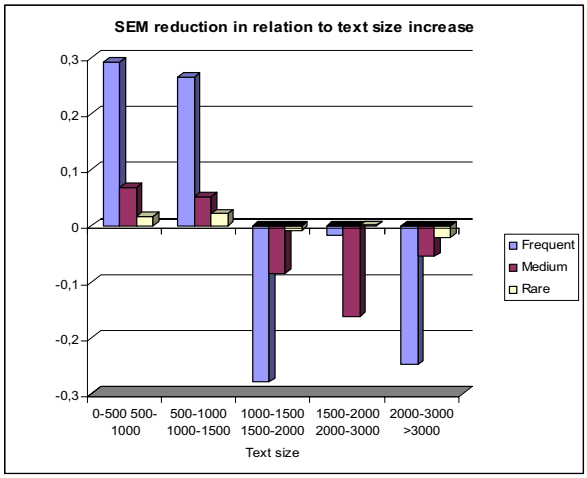
By examining the graph 4 we note that the text size interacts with the frequency of a feature. The frequent

features exhibit curvilinear correlation with the text size and their SEM decreases with high rate up to the text size of 1500 words. However, the medium frequency and the rare features do not correspond to the text size increase. They have flat response all over the text size continuum.

The investigation of the SD gave similar results with the SEM evolution and is displayed in the graph 5. The frequent features are very sensitive to the text size increase and their SD falls sharply as the text size increases. The largest reduction in SD once again it occurs in the transition from the 0 - 500 to 500 - 1000 words. On the other hand, the medium frequency and the rare features do not correlate to the text size and its increase does not affect significantly their SD. The effect of the text size increase in the SD and SEM reduction is shown in the graph 6 & 7.



Graph 6: SD reduction rate in relation to text size increase



Graph 7: SEM reduction rate in relation to text size increase

In both graphs it is evident that:

a) For the frequent and medium frequency linguistic features the largest reduction is taking place in the transition from the 0-500 to 500-1000 text size frame.

b) For the rare features even large augmentations in the text size do not induce significant reduction in their SD and SEM values.

4. Conclusions

The present study tried to approach the issue of the estimation of the optimum sample size for Modern Greek corpora adopting a user oriented view. We constructed a test corpus based on augmenting text size frames and we searched in each one a number of linguistic features that belonged to a variety of linguistic levels. The concept of optimal text size in this test corpus was approached through the investigation of 3 quantitative parameters: (PNA, SD, SEM)

These parameters were studied in relation to the text size and the linguistic feature overall frequency. The results indicate that:

- The most influential factor regarding the feature's occurrence in a corpus is its overall frequency.
- The linguistic level of the features does not affect their representativity in text files of different sizes.
- The appearance of medium frequency features is strongly correlated with the size of the text files that comprise a corpus.
- The highest reduction of the SD and SEM of the features was observed in the increase of the text size frame from 0 - 500 to 500 – 1000.
- The SD and SEM of the rare features are not influenced from the text size increase.

5. References

Biber, D. & Finegan E. (1991). On the exploitation of computerized corpora in variation studies. In K. Aijmer & B. Altenberg (Eds.), English corpus linguistics (pp. 204 – 220). London: Longman.

de Haan, P. (1992). The optimum corpus sample size?. In G. Leitner (Ed.), New dimensions in English language corpora (pp. 3 – 19). Berlin: Mouton de Gruyter.

EAGLES, (1996). Preliminary recommendations on Corpus Typology. EAGLES Document: EAG-TCWG-CTYP/P.

Kucera, H. & Frances, W. N. (1967). Computational analysis of present-day American English. Providence. RI: Brown University Press.

Hatzigeorgiu N., Gavrilidou M., Piperidis S., Carayannis G., Papakostopoulou A., Spiliotopoulou A., Vacalopoulou A., Labropoulou P., Mantzari E., Papageorgiou H., Demiros I. (2000). Design and implementation of the online ILSP Greek Corpus. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhaouer (Eds.), Proceedings of the Second International Conference on Language Resources and Evaluation (pp. 1737-1742). Athens.