Journal of Quantitative Linguistics
2005, Vol. 12, No. 2-3, pp. 167 – 184
DOI: 10.1080/09296170500172478

Routledge
Taylor & Francis Group

Basic Quantitative Characteristics of the Modern Greek Language Using the Hellenic National Corpus*

George Mikros[1,2], Nick Hatzigeorgiu[2], & George Carayannis[2]
[1]Department of Italian and Spanish Language and Literature, School of Philosophy, University of Athens [2]Institute for Language and Speech Processing (ILSP)

ABSTRACT

Modern Greek is one of the least quantitatively studied modern European languages and the goal of this paper is to fill this relative void. We use the Hellenic National Corpus (HNC), which is a growing corpus that currently includes 33 million words. The corpus and all the tools used in our work were developed by the Institute for Language and Speech Processing (ILSP). In this paper we focus on three main areas: the lists of the 1000 most common words and lemmas, word length and letter frequency. We also make some comparisons with earlier work, in which we had used the previous 13 million word edition of the HNC.

INTRODUCTION

The quantitative study of the structure of language has become one of the most important goals of current linguistic research. The high availability of electronic corpora, combined with the development of the required computing and statistical techniques for the handling of language data, has introduced quantitative methods in the analysis of every linguistic aspect (Bod et al., 2003). The increased use of quantitative methods in linguistic studies has also received a boost from the growing success of these methods in the field of natural language processing (Manning & Schütze, 1999). This general trend can also be observed in the linguistic research on the Modern Greek (MG) language. In a recent study (Mikros, in press), we found that the use of quantitative methods in the linguistic

*Address correspondence to: Dr. George K. Mikros, Department of Italian and Spanish Language and Literature, School of Philosophy, University of Athens, Panepistimioupoli Zografou, GR-157 84, Athens, Greece. E-mail: gmikros@isll.uoa.gr

research of MG has followed an exponential growth, and the percentage of quantitative studies of MG in the 1990s is five times that of the 1980s.

An important development for the quantitative research of MG is the development of the Hellenic National Corpus (HNC) (Hatzigeorgiu et al., 2000) by the Institute for Language and Speech Processing (ILSP). In our earlier work (Hatzigeorgiu et al., 2001) we used the first edition of the HNC (13 million words) to study the applicability of Zipf's law on the 1000 most common words and lemmas. We had also computed some basic quantities, such as the average word length and the distribution of the grammatical categories. That work gave the first quantitative results for high frequency words. Since then, HNC has grown to 33 million words, which renders it interesting to study how those quantities have changed. In particular, the goals of this paper are:

- Comparison of the 1000 most common words and lemmas for the two editions of the HNC.
- Investigation of the applicability of Zipf's law for the new edition of HNC.
- Publication of the first frequency list for the letters of MG based on HNC.
- Investigation of the distribution of the word length in MG.

We believe that this work will assist the studies of the automatic document categorization which are under way (Mikros & Carayannis, 2000; Tambouratzis et al., 2000) and will also help linguistic work on MG, such as the development of intelligent search engines for Greek Web pages. We note that up to now little work has been completed on quantitative studies of the MG language, compared with other European languages (e.g., Saukkonen, 1994; Hammerl & Sambor, 1993; Tešitelová, 1992).

THE HNC CORPUS

The corpus used for the investigations of this paper is the HNC. The HNC, all the tools used for its development, as well as all the tools used in this paper, were developed by the ILSP. The HNC is an ongoing effort.[1] It currently contains more than 48,000 written MG texts, published from 1976 on, totalling 33 million words.

[1] The HNC has a Web interface and queries are possible over the Internet at the following Web address: http://hnc.ilsp.gr/

Table 1. HNC Text Genre Distribution.

Medium	Percentage of words
Book	10
Newspaper	79
Periodical	4.5
Miscellaneous	6.5

Texts in HNC are classified according to PAROLE standards (PAROLE, 1995), which follow the TEI (Sperberg-McQueen & Burnard, 1994) and EAGLES (EAGLES, 1994) guidelines. Texts are classified with regards to medium, genre, topic, detailed genre, detailed topic and bibliographical information. As far as medium is concerned, texts are classified into four categories, according to their source.[2] The current percentage of words for each one of the four categories can be seen in Table 1.

The HNC is a growing corpus. Its size is changing as new documents are added to the existing ones. In this paper we use the latest edition that contains 33 million words. We denote this particular incarnation of the HNC as the 33MW HNC. Sometimes we make comparisons with results from an earlier version of the HNC, the 13MW HNC.

QUANTITATIVE CHARACTERISTICS OF THE 1000 MOST COMMON WORDS AND LEMMAS

Comparisons for the 1000 Most Common Words

Hatzigeorgiu et al. (2001) published the first list of the 100 most common words and lemmas in MG, using the first edition of HNC, which contained 13 million words (MW). Now that HNC has grown to 33 MW, it is important to review those findings and to compare them with the new data.

We had found that the 1000 most common words accounted for the 59.9% of the 13MW HNC. In the new 33MW HNC, the 1000 most common words account for the 60.4% of the corpus. We conclude,

[2] *A propos*, we would like to thank all the publishers that have donated the texts used in HNC.

therefore, that the 1000 most common words in each version of HNC account for a constant percentage of the whole corpus, even though the corpus grew by a factor of 2.6.

Next, we performed a more detailed examination of the differentiation of the most common words in the two versions of the HNC. The two lists with the 1000 most common words for each version of the HNC contain rank and frequency information (as a percentage of the total words contained in the corpus). The two lists contain 895 common words, i.e., 89.5% of the words are present in both lists.

We decided to use the non-parametric Wilcoxon signed rank test[3] in order to find out if there is a significant statistical difference between the two lists. The results from the Wilcoxon test show that there is no significant statistical differentiation between the two lists, either for the relative ranking of the 1000 most common words ($z = -0.808$, $p = 0.41$), or for their relative frequency ($z = -1.721$, $p = 0.08$). Consequently, while HNC has grown, there are no significant statistical differences for the 1000 most common words of the corpus.

To verify this result, we also studied the correlation of the two lists, for both the ranking and the frequency, using the Spearman r (r_s) statistic. Using the Spearman rank correlation for the relative ranks of the words in the two lists we got $r_s = 0.89$, $p < 0.001$, while for the relative frequencies we got $r_s = 0.90$, $p < 0.001$. These results show a high correlation for the two lists.

The scatter plot (Fig. 1) for the relative ranking for the words in the two lists also shows that the ranking of the common words remains almost constant for the two versions of the HNC.

While the comparison of the 1000 most common words for the two versions of the HNC shows that there is little difference between them, a similar comparison of the 1000 most common lemmas for the two versions of the HNC shows far greater differences. Common lemmas in the two lists are only 775 (77.5%). Also, the Wilcoxon Signed Rank Test shows a significant differentiation, both for the ranking of lemmas ($z = -4.64$, $p < 0.001$) and for the relative frequency of lemmas ($z = -2.16$, $p < 0.05$).

[3] This test is preferable to other statistical tests, since it does not require a specific distribution for the test parameters. The Wilcoxon test uses the differences between two pairs of measurements and gives higher weight to pairs that have a greater difference.

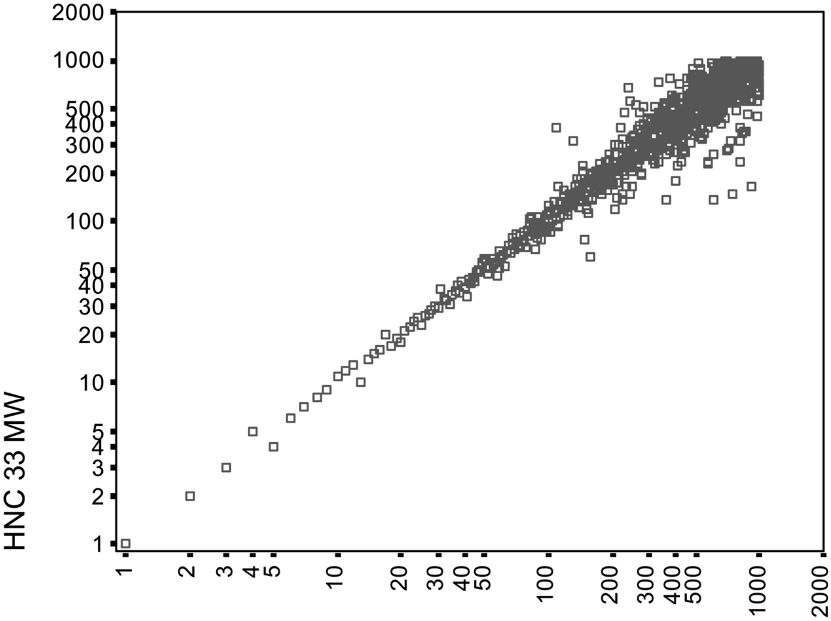

Fig. 1. Log scatter plot displaying the correlation of the 1000 most common words for the two editions of the HNC (13MW and 33MW)

Zipf's Law for Words and Lemmas

One of the most well-known laws in quantitative linguistics is Zipf's law that connects the rank of the word to its frequency. While there are some remarks by the French psychologist Estoup (Tesitelova, 1992, p. 50) at the end of the 19th century, it was Zipf that made this connection well known. In particular, Zipf's first law is a power law, which states that the rank (r_i) of a member of an ordered list is connected to the frequency of appearance of this member (p_i) by the following equation:

$$p_i = \frac{b}{r_i^a} \Rightarrow \log(p_i) = B - a\log(r_i), \text{ with}\, a \approx 1 \tag{1}$$

The validity of this empirical law has been observed in a large spectrum of phenomena, including natural languages, economics, biological systems and even in statistics of Web usage.

The validity of Zipf's first law has been verified for a large number of languages (Miller et al., 1958; Rousseau & Zhang, 1992). An explanation for this law is that the language system strives to balance the frequency of appearance of a word with the number of words that share the same frequency of appearance. This balancing trend is a result of two opposing forces. The first tries to limit lexical plurality, and this force could lead a language to have a minimum number of words with the maximum number of appearances for each. The opposing force tries to maximize lexical plurality in order to achieve maximal clarity, and the theoretical limit would be the maximum number of words with the minimum number of occurrences. These two forces correspond to the important aspects of any kind of communication. The transmitter wants to code the message using the least possible effort and the least possible amount of words. On the contrary, the receiver prefers the maximum possible information contained in the message, so that a minimal effort will be required to decipher the message.

In HNC, the frequency of occurrence for the 1000 most common words follows Zipf's law quite closely, both in the 13MW version and in the 33MW version. Figure 2 shows the corresponding diagrams. For the 13MW version of HNC, parameter α of equation (1) has a value of 0.96, while for the 33MW version its value is 0.97. Neither value is accurate for the initial 20 points of the diagram that do not lie on a straight line. It is also apparent that the parameter a has almost the same value for both versions of HNC and that its value is very close to 1, as expected. Finally, the curves have quite similar shapes, even thought the size of the corpus has grown considerably.

The investigation of Zipf's law for the lemmas leads to quite similar results. Of course, in this case we also have some errors due to the way we determine the lemma for each word. However, we can see in Figure 3 that these errors have little influence on Zipf's law, except that the first few words deviate from the expected straight line. The parameter a of equation (1) is 0.87 for the 13MW HNC and 0.90 for the 33MW HNC.

WORD LENGTH IN MODERN GREEK

Word lengths and distributions have been studied quite extensively in quantitative linguistics. One recent example is the Göttingen project (Best, 1998). For the Indo-European languages, the distribution of the

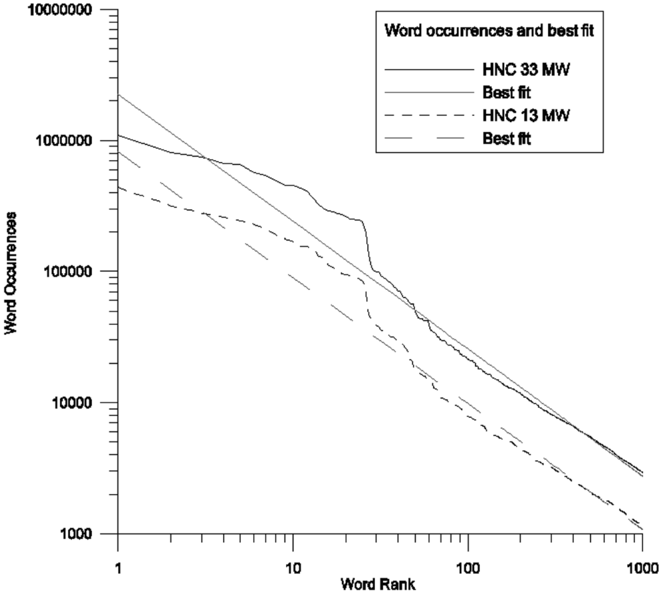

Fig. 2. Zipf's first law for words. Occurrences of words in logarithmic scale for the two versions of HNC.

word lengths has been studied since the middle of the 20th century, with the pioneering work of the Russian mathematician Čebanov (Altmann, 1988, p. 58). Fucks (1956), in one of the first comparative studies on the distribution of word lengths, investigated eight languages and found that they follow the "1 displaced poisson" distribution. More recent studies (Grotjahn, 1982, p. 68) have shown that the "negative binomial" distribution is more appropriate, since it does not assume that the probabilities for single words are equal, but takes into consideration the dependence on concordances and other factors. Altman (1988, p. 58) attempted to model word length distributions connecting the opposing forces of the language when it is viewed as information, in a way similar to the explanation of Zipf's first law (see section above). The comparative analysis (Best, 1998, p. 158) of 38 languages that belong to all major language families (including Ancient Greek) has shown that the "hyper-poisson" distribution is a satisfactory model for word lengths.

Even though the average word length in a text is important for the determination of the text style, and has been already used in text style

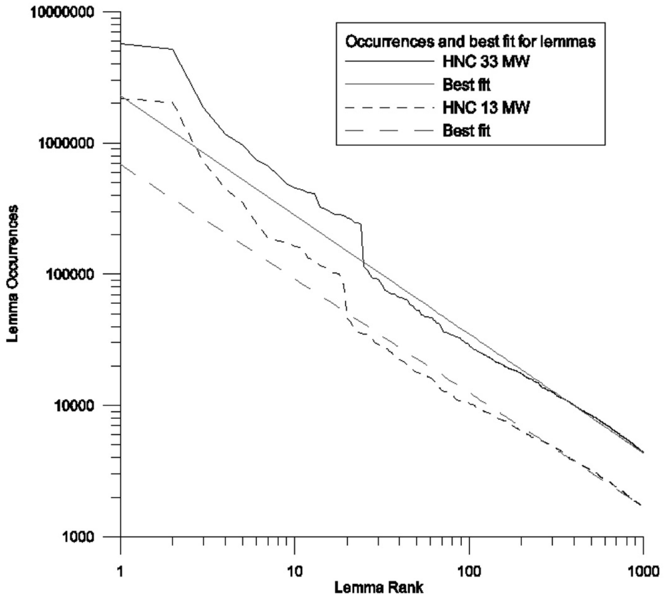

Fig. 3. Zipf's first law for lemmas. Occurrences of lemmas in logarithmic scale for the two versions of HNC.

studies (Bekiari et al., 2001) and in automatic text categorization research (Mikros & Carayannis, 2000; Tambouratzis et al., 2000), there are no systematic studies of the word length distribution in Modern Greek. In the following, we will examine the word lengths for the HNC and its 1000 most common words. Furthermore, we will study the distribution of word lengths in some Modern Greek texts and we will compare our findings with similar studies of other languages.

Word Length in HNC

Average word length in HNC is 5.33 letters. In general, this word length is not homogeneous. Word length is a quantity that depends on many factors, including the publication medium (Wimmer et al., 1994, p. 99). The influence of publication medium on the average word length of HNC texts can be seen in Table 2.

Figure 4 shows that in HNC the diagrams for word lengths for each of the four publication categories are quite similar. There are small differences only for the words with medium lengths (4 – 10 letters).

Table 2: Average Word Length for each Publication Medium of HNC.

HNC	Books	Newspapers	Periodicals	Miscellaneous
5.33	5.41	5.29	5.38	5.56

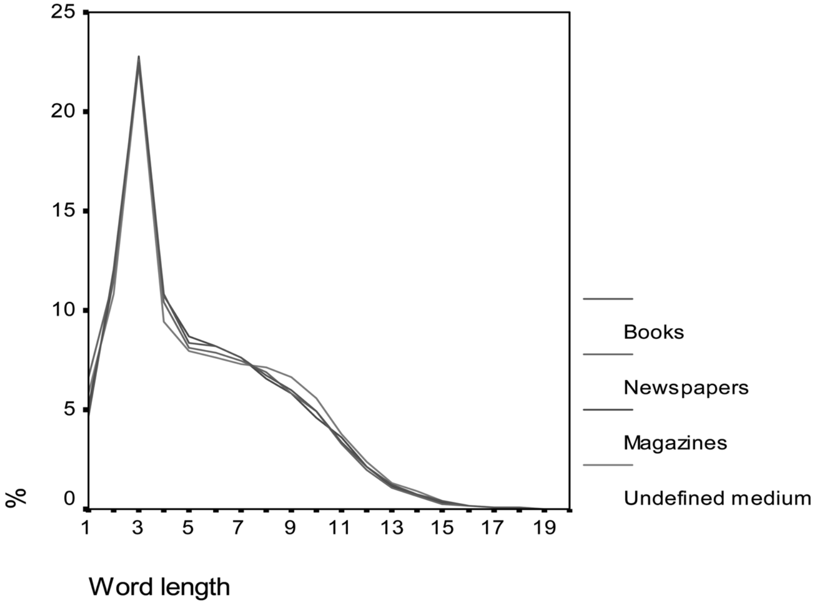

Fig. 4. Word length distribution in HNC for each of the four publication media.

Also, when we compare the distributions for the word length of the 1000 most common words with the distribution for the word lengths of the whole corpus (Fig. 5), we see that the 1000 most common words show a higher concentration in shorter words (1 – 5 letters), while there is a larger area covered by the distribution of words for the whole corpus.

It is also interesting to examine how the average word length changes for the most common words. If α_i is the length of the word i and b_i is the number of appearances of the word i, then we define the running average for the word lengths as:

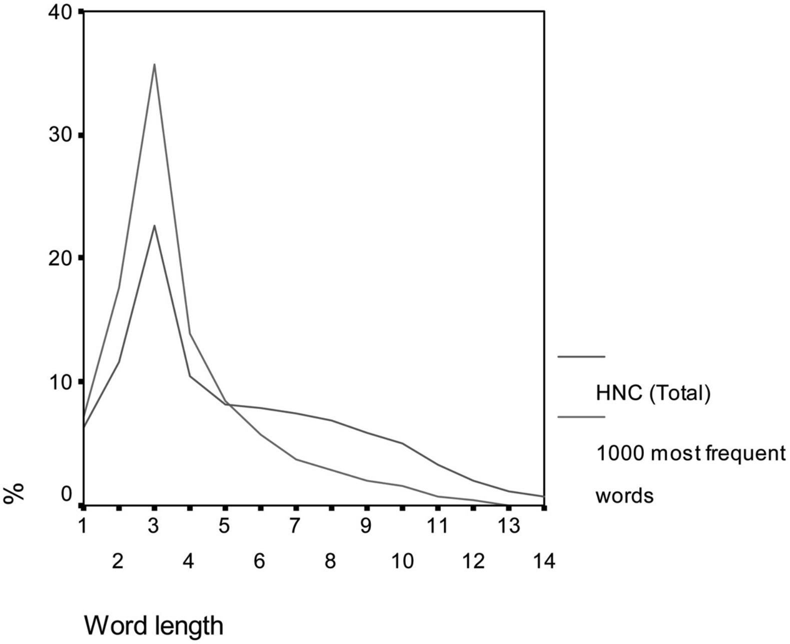

Fig. 5. Comparison of word lengths for the 1000 most common words and for the whole HNC.

$$y_i = \frac{\sum_0^i a_i \cdot b_i}{\sum_0^i b_i}.$$

As we can see in Figure 6, the running average word length increases monotonically until it becomes equal to the total average word length of 5.33 letters (straight line). It is also obvious that the most common words tend to be shorter than the average word length of the HNC. Both of these observations comply with what has been found for other languages (Grotjahn & Altmann, 1993) and they are compatible with Zipf's "principle of least effort" and the more general self-organization of language systems, which is apparent in many aspects of their structure.

We note that the above results for the average word lengths and the corresponding distributions are almost identical with those that we made

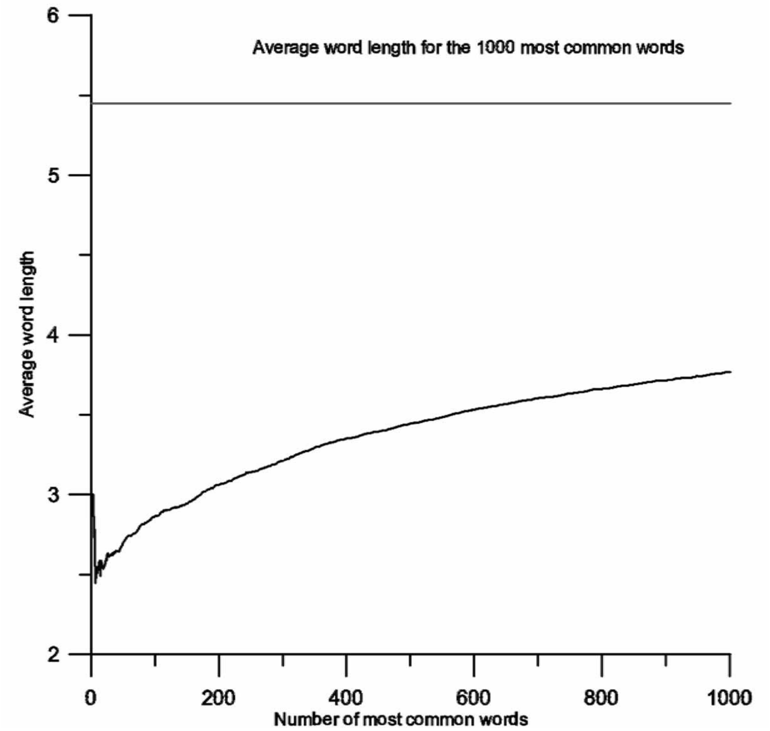

Fig. 6. Running average word length for the 1000 most common words.

previously (Hatzigeorgiu et al., 2001) for the 13MW HNC. The only difference is the decrease of the average word length, from 5.45 letters in the 13MW HNC to 5.33 letters in the 33MW HNC. This small decrease can be explained by the fact that the distribution of publication media has changed, and newspapers now constitute a larger percentage of the corpus. As we have seen in Table 2, newspapers have shorter average word length, so it is expected that the average word length for the whole corpus will decrease as well.

In addition to the above analysis, we decided to investigate word length in single documents. To do this, we randomly selected six books of various subjects and we computed their distributions of word lengths. We found that the distributions follow quite closely the negative binomial distribution, which is given by the following equation:

$$f(x) = \binom{s+x-1}{x} p^s (1-p)^x,$$

where s = number of successes, $s > 0$, and p = probability of a success, $0 < p < 1$.

The results of the best fit of the data to the negative distribution can be seen in Table 3.

This distribution has been used for the fit of other languages as well (Best, 1998, p. 157; Grotjahn, 1988, p. 55; Wimmer & Altmann, 1996), even though many European languages seem to follow the Hyper-Poisson distribution (Best, 1998, p. 158). We plan to make a more detailed and a more complete investigation of the word length distributions in MG.

LETTER FREQUENCIES

The total number of Greek characters in HNC is 166,644,226. The specific number of appearances for each letter can be seen in Table 4. Each letter in these results contains both non-accented and accented letters, capitals and lower-case letters. The corresponding distribution can be seen in Figure 7.

Furthermore, we investigated the distribution of accented and non-accented letters and the results are presented in Table 5. Accented vowels constitute the 22% of the total number of vowels, and their percentage for each vowel varies from about 33% for the Ω to 16% for the A.

Table 3. Results for the Best Fit for Word-Length Data from Six Documents of HNC to the Negative Binomial Distribution.

	Distribution parameters		Best fit	
Document type	s	P	χ^2	$p(\chi^2)$
Scientific study 1	1	2.3E -4	4.41	0.21
Scientific study 2	1	5.6E -4	5.16	0.16
Scientific study 3	1	2.4E -4	3.00	0.39
Novel 1	1	6.6E -4	4.83	0.18
Novel 2	1	1.04E -4	4.82	0.18
Law document	1	1.9E -4	5.16	0.15

Table 4. Frequencies for Each Greek Letter in HNC.

Letter	Appearances	Percentage	Letter	Appearances	Percentage
Α	19,016,573	11.411	Μ	5,596,727	3.358
Ο	17,216,358	10.331	Λ	4,553,342	2.732
Ι	15,417,480	9.252	Ω	3,577,468	2.147
Ε	14,308,090	8.586	Δ	2,915,228	1.749
Τ	13,194,359	7.918	Γ	2,878,003	1.727
Σ	13,047,535	7.830	Χ	1,962,412	1.178
Ν	10,329,547	6.199	Θ	1,866,867	1.120
Η	8,997,302	5.399	Φ	1,352,884	0.812
Υ	7,359,535	4.416	Β	1,136,565	0.682
Ρ	7,141,941	4.286	Ξ	669,155	0.402
Π	6,689,110	4.014	Ζ	574,177	0.345
Κ	6,622,486	3.974	Ψ	221,082	0.133

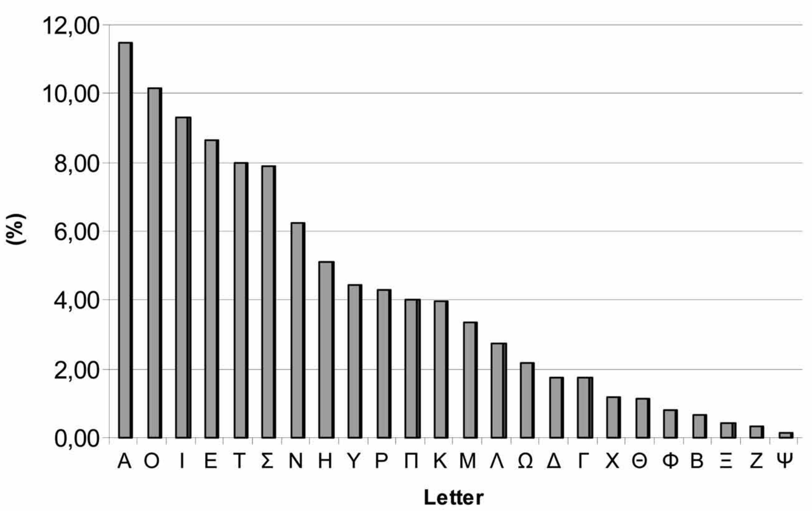

Fig. 7. Relative frequency of letters in MG, according to their appearance in HNC.

We have also investigated the frequency for each letter as a function of their position in a word (beginning, end, or middle of word). The results are shown in Table 6 and the corresponding diagram can be seen in Figure 8.

Table 5. Distribution of Accented and Non-Accented Greek Letters in HNC.

Accented	Frequency	%	Non-Accented	Frequency	%	Total
ά	3,310,673	17.43	α	15,680,065	82.57	18,990,738
έ	3,064,543	21.45	ε	11,223,397	78.55	14,287,940
ό	3,503,641	20.91	ό	13,252,900	79.09	16,756,541
ώ	1,204,832	33.69	ω	2,371,517	66.31	3,576,349
ί, ϊ, ΐ	4,033,474	26.17	ι	11,379,680	73.83	15,413,154
ή	2,207,937	26.27	η	6,196,292	73.73	8,404,229
ύ, ϋ, ΰ	1,658,234	22.54	υ	5,699,914	77.46	7,358,148

Table 6: Appearances of each letter and their relative frequency with respect to their position in a word

	Initial	%	Middle	%	End	%	Total
Α	3,027,599	15.92	10,814,960	56.87	5,174,014	27.21	19,016,573
Β	406,797	35.79	710,905	62.55	18,863	1.66	1,136,565
Γ	889,836	30.92	1,965,217	68.28	22,950	0.80	2,878,003
Δ	1,434,827	49.22	1,451,962	49.81	28,439	0.98	2,915,228
Ε	3,055,110	21.35	9,385,872	65.60	1,867,108	13.05	14,308,090
Ζ	99,660	17.36	462,181	80.49	12,336	2.15	574,177
Η	826,501	9.19	5,051,230	56.14	3,119,571	34.67	8,997,302
Θ	533,658	28.59	1,318,805	70.64	14,404	0.77	1,866,867
Ι	291,969	1.89	10,924,053	70.85	4,201,458	27.25	15,417,480
Κ	2,648,806	40.00	3,803,065	57.43	170,615	2.58	6,622,486
Λ	336,735	7.40	4,148,222	91.10	68,385	1.50	4,553,342
Μ	1,940,285	34.67	3,605,974	64.43	50,468	0.90	5,596,727
Ν	960,140	9.30	5,578,176	54.00	3,791,231	36.70	10,329,547
Ξ	97,548	14.58	562,470	84.06	9,137	1.37	669,155
Ο	1,934,754	11.24	11,704,050	67.98	3,577,554	20.78	17,216,358
Π	2,970,971	44.42	3,650,980	54.58	67,159	1.00	6,689,110
Ρ	141,042	1.97	6,909,425	96.74	91,474	1.28	7,141,941
Σ	2,580,472	19.78	4,774,831	36.60	5,692,232	43.63	13,047,535
Τ	4,921,137	37.30	8,173,431	61.95	99,791	0.76	13,194,359
Υ	407,969	5.54	4,832,634	65.66	2,118,932	28.79	7,359,535
Φ	334,348	24.71	1,000,590	73.96	17,946	1.33	1,352,884
Χ	490,047	24.97	1,437,958	73.28	34,407	1.75	1,962,412
Ψ	44,054	19.93	176,695	79.92	333	0.15	221,082
Ω	119,559	3.34	3,177,579	88.82	280,330	7.84	3,577,468

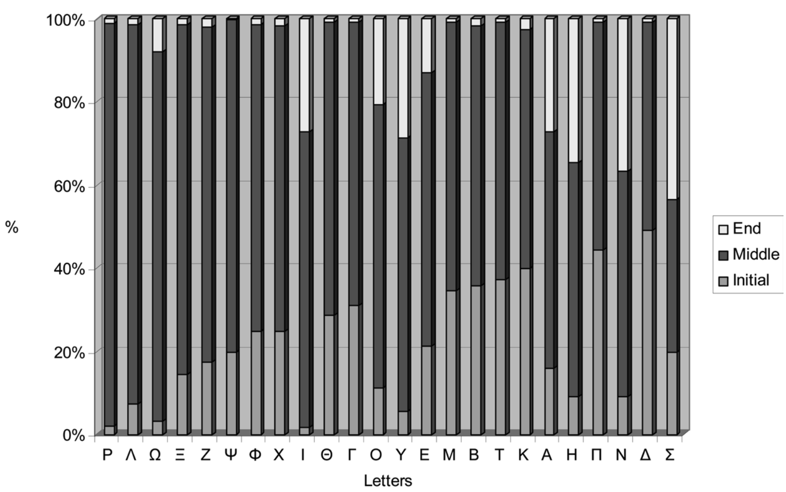

Fig. 8. Comparison diagram for the positions of each letter in a word.

It is evident from the data that letters Δ, Π, Κ and Τ are the most common consonants at the beginning of a word, while Ε and Α are the most common vowels. At the end of a word, the most common consonants are Σ and Ν, and the most common vowels are Η and Υ. Finally, consonants Π and Λ are almost exclusively present in the middle of words (probabilities are 97% and 91% respectively).

In a recent study for the appearance of the space character in many European languages, Rosenbaum & Fleischmann (2002, p. 242) argue that MG has a higher appearance of space characters than the Romance languages. In particular, they claim that space in MG appears with a frequency of 19.4% in the corpora that they have used, while in Latin and in Romance languages the frequency is 14.6%. However, the HNC data show that the space character has exactly the same frequency of appearance as the Romance languages, i.e., 14.6% of the total characters of HNC. Our results for the appearances of numerical characters in the HNC (Table 7) are also different from those in (Rosenbaum & Fleischmann, 2002).

Differences exist also for the combined letter frequencies reported in Rosenbaum & Fleischmann (2003, pp. 35–36). For a detailed comparison of these figures in the Rosenbaum & Fleischmann study and the HNC see Table 8.

Table 7. Comparison of the Relative Frequency of Numerical Digits for the Corpus of Rosenbaum & Fleischmann and for the HNC.

Numerical character	Rosenbaum & Fleischmann	%	HNC	%
0	10,041	13.3	308,094	17.4
1	15,850	20.9	344,033	19.5
2	8,106	10.7	184,785	10.4
3	6,199	8.2	129,854	7.3
4	5,331	7.0	106,680	6.0
5	5,606	7.4	131,885	7.5
6	5,276	7.0	104,937	5.9
7	5,198	6.9	108,630	6.1
8	6,559	8.7	99,651	5.6
9	7,543	10.0	250,125	14.1

Table 8. Comparison of the Letter Frequencies and their Relative Frequencies for the Corpus of Rosenbaum & Fleischmann and for the HNC.

Letters	HNC	%	Rosenbaum & Fleischmann	%
Α	19,016,573	11.411	927,520	12.334
Ο	17,216,358	10.331	769,631	10.234
Ι	15,417,480	9.252	700,746	9.318
Ε	14,308,090	8.586	660,292	8.780
Τ	13,194,359	7.918	600,534	7.986
Σ	13,047,535	7.830	564,325	7.504
Ν	10,329,547	6.199	484,208	6.439
Η	8,997,302	5.399	340,993	4.534
ϒ	7,359,535	4.416	331,957	4.414
Ρ	7,141,941	4.286	305,317	4.060
Π	6,689,110	4.014	306,324	4.073
Κ	6,622,486	3.974	283,847	3.775
Μ	5,596,727	3.358	267,339	3.555
Λ	4,553,342	2.732	193,850	2.578
Ω	3,577,468	2.147	157,894	2.100
Δ	2,915,228	1.749	133,525	1.776
Γ	2,878,003	1.727	126,724	1.685
Χ	1,962,412	1.178	89,688	1.193
Θ	1,866,867	1.120	84,423	1.123
Φ	1,352,884	0.812	66,676	0.887
Β	1,136,565	0.682	47,994	0.638
Ξ	669,155	0.402	35,326	0.470
Ζ	574,177	0.345	30,147	0.401
Ψ	221,082	0.133	10,831	0.144

CONCLUSIONS

In this work, we have presented some basic quantitative characteristics of the Modern Greek language, using the 33MW HNC, developed by the ILSP. The HNC is a growing corpus and its size is changing, but we have found that the list of the 1000 most common words is stable as the size of the corpus grows. Also, Zipf's law is valid for the 33MW HNC as it was for the 13MW HNC and the shapes of the corresponding curves do not change much as the corpus grows.

The average word length depends on many factors and in particular the publication medium. The "newspaper" category of the HNC has an average word length of 5.29 letters per word and the "miscellaneous" category 5.56 letters per word. Newspaper articles are dominant in the corpus and as a result of this, the average word length for the HNC is 5.33 letters per word.

We have also presented some analytic results for the frequencies of the letters in MG. One interesting conclusion of our investigation is that the frequency of the space character in the MG texts of HNC is 14.6% of the total characters of HNC, the same frequency found for Latin and Romance languages.

This work is part of our ongoing effort for a quantitative study of the MG language, which, as pointed earlier, is one of the least studied modern European languages. We hope that our work will be the prelude to a more detailed quantitative study of MG.

REFERENCES

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Bekiari, Ch., Papavasileiou, V., & Pasxalis S. (2001). Literary stylometry with applications to authorship attribution [in Greek]. Unpublished MA Thesis, Interdepartmental Postgraduate Program "Technoglossia", Athens.

Best, K-H. (1998). Results and perspectives of the Göttingen project on quantitative linguistics. *Journal of Quantitative Linguistics*, *5*, 155–162.

Bod, R., Hay, J., & Jannedy, S. (Eds) (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

EAGLES. (1994). *Corpus encoding: Draft*. Technical report, EAGLES. Document EAG-CSG/IR-T21.

Fucks, W. (1956). Die mathematischen Gesetze der Bildung von Sprachelementen aus ihren Bestandteilen. *Nachrichtentechnische Fachberichte*, *3*, 7–21.

Grotjahn, R. (1982). Ein statistisches Modell zur Verteilung der Wortlänge. *Zeitschrift für Sprachwisswenschaft*, *1*, 44–75.

Grotjahn, R., & Altmann, G. (1993). Modelling the distribution of word length. In R. Köhler & B. B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 141–153). Dordrecht: Kluwer.

Hammerl, R., & Sambor, J. (1993). Synergetic studies in Polish. In R. Köhler & B. B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 331–359). Dordrecht: Kluwer.

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 et al. (Eds.), *Proceedings of the LREC 2000 Conference* (pp. 1737–1742). Athens.

Hatzigeorgiu, N., Mikros, G., & Carayannis, G. (2001). Word length, word frequencies, and Zipf's law in the Greek language. *Journal of Quantitative Linguistics*, *8*, 175–185.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mikros, G. (in press). Quantitative linguistics in Greece. In G. Altmann, R. Köhler, & R. Piotrowski (Eds.), *Quantitative Linguistics. An international handbook*. Berlin: Walter De Gruyter.

Mikros, G., & Carayannis, G. (2000). Modern Greek Corpus Taxonomy. In M. Gavrilidou et al. (Eds.), *Proceedings of the LREC 2000 Conference* (pp. 129–134). Athens.

Miller, G. A., Newman, E. B., & Friedman E. A. (1958). Length-frequency statistics for written English. *Information and Control*, *1*, 370–389.

Rosenbaum, R., & Fleischmann, M. (2002). Character frequency in multilingual corpus 1 – Part 1. *Journal of Quantitative Linguistics*, *9*, 233–260.

Rosenbaum, R., & Fleischmann, M. (2003). Character frequency in multilingual corpus 1 – Part 2. *Journal of Quantitative Linguistics*, *10*, 1–39.

Rousseau, R., & Zhang, Q. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics*, *24*, 201–220.

Saukkonen, P. (1994). Main trends and results of quantitative linguistics in Finland. *Journal of Quantitative Linguistics*, *1*, 2–15.

Sperberg-McQueen C. M., & Burnard, L. (1994). *Guide-lines for electronic text encoding and interchange: Tei-p3*. Technical report, Chicago and Oxford, ACH-ACL-ALLC Text Coding Initiative.

Tambouratzis G., Markantonatou, S.., Hairetakis, N., & Carayannis, G. (2000). Automatic style categorization of corpora in the Greek language. In M. Gavrilidou et al. (Eds.), *Proceedings of the LREC 2000 Conference* (pp. 135–140). Athens.

Tešitelová, M. (1992). *Quantitative linguistics*. Amsterdam & Philadelphia: John Benjamins.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, *3*, 38–50.

Wimmer, G., & Altmann, G. (1996). The theory of word length: some results and generalizations. *Glottometrika*, *15*, 112–133.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length. *Journal of Quantitative Linguistics*, *1*, 98–106.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA.: Addison Wesley.