Extracting Collocations in Modern Greek Language

Kostas Fragos[1], Yannis Maistros[2], Christos Skourlas[3]

1 Department of Computer Engineering, National technical University of Athens, Iroon Polytexneiou 9 15780 Zografou Athens Greece
http://nts.ece.ntua.gr/nlp_lab
kfragos@ece.ntua.gr
2 Department of Computer Engineering, National technical University of Athens, Iroon Polytexneiou 9 15780 Zografou Athens Greece
maistros@ece.ntua.gr
3 Department of Computer Science, Technical Educational Institute of Athens, Ag Spyridonos 12210 Aigaleo Athens Greece
cskourlas@teiath.gr

Abstract. In this paper we describe and apply two statistical methods for extracting collocations from text corpora written in Modern Greek. The first one is the *mean and variance* method which calculates "offsets" (distances) between words in a corpus and looks for patterns of distances with low spread. The second method is based on the X^2 test. Such an approach seems to be more flexible because it does not assume normally distributed probabilities of the words in the corpus. The two techniques produce interesting collocations that are useful in various applications e.g. computational lexicography, language generation and machine translation.

1 Introduction

Collocations are common in Natural Languages and can be met in technical and non-technical texts. A collocation could be seen as a combination of words (or phrases) which are frequently used together in a way that sounds correctly. Collocations in Natural Languages with rich inflectional system (e.g. Modern Greek) could also be seen as phrases where nouns appear under one "rigid" form, that is, they appear with the same syntactic way in all their occurrences, e.g. the Greek words "*Χρηματιστήριο" and "Αξιών"* are only combined in the collocation "*Χρηματιστήριο Αξιών*" (*Stock Exchange*). Other phrases are more "flexible" e.g. the Greek words "Σ*τρώνω / στρωνομαι"* and *"δουλειά"* could be combined in various phrases having different meaning, as the following ones:

"Στρώνομαι στην δουλειά" (*To get to work*)
"Η δουλειά μου στρώνει" (*My business is looking up*).

There are different definitions because many researchers have focused on different aspects of collocations. According to Firth [6], "Collocations of a given word are statements of the habitual or customary places of the word".

Benson and Morton [1] define collocations *as an arbitrary and recurrent word combination*. The word *recurrent* means that these combinations are common in a

given context. Smadja [15] identifies four characteristics of collocations useful for machine applications:
a) Collocations are arbitrary; this means that they do not correspond to any syntactic or semantic variation. b) Collocations are domain-dependent; hence handling text in a domain requires knowledge of the related terminology / terms and the domain-dependent collocations. c) Collocations are recurrent (see above) d) Collocations are cohesive lexical clusters; by cohesive lexical clusters is meant that the presence of one or several words often implies or suggests the rest of the collocation.
In the work of Lin [10] collocation is defined as a habitual word combination. Gitsaki et. al. [7] define it as a recurrent word combination. Howarth and Nesi [8] have approached the use of collocations from the foreign language learner perspective.

According to Manning and Schutze [11] collocations are characterized by *limited compositionality*. A natural language expression is compositional if the meaning of the expression can be predicted from the meaning of the parts. Hence, collocations are not fully compositional. For example in the Greek expression "*γερό ποτήρι*" (*heavy drinker*), the combination has an extra meaning, a person who drinks. This meaning is completely different form the meaning of the two collocates (portions of the collocation): "*γερό"* (*strong*), "*ποτήρι"* (*glass*). Another characteristic of collocations is the lack of valid synonyms for any collocates [11], [10]. For example, even though *baggage* and *luggage* are synonyms we could only write *emotional*, *historical* or *psychological baggage.*

2 The Rationale for Extracting Collocations in NLP Applications

Collocations are important for a number of applications e.g. Natural Language generation, machine translation, text simplification and computational lexicography.
Howarth and Nasi [8] claimed that most natural language expression contain at least one collocation. Natural language generation requires knowledge about the valid combinations of the words.

Machine Translation (MT) is considered as one of the most difficult tasks in natural language processing, and in artificial intelligence in general. Accurate translation seems to be impossible without the comprehension of the text. It is a difficult task to translate Collocations across languages. This fact has a direct implication in machine translation applications. According to Gitsaki [7] collocations differ from language to language. For example a *clear road* in English is "ελεύθερος δρόμος" (*free road)* in Greek.

Collocational information is also crucial in text simplification tasks. This involves techniques of replacing difficult words with simpler ones. Without the knowledge of collocations and the related constraints, this replacement can lead to awkward text. An example of simplification tasks is the *Practical Simplification of English Text* (PSET) project [2].
Collocations are also important in computational lexicography. They are used to fully characterize the lexical entries. According to Richardson "For a detailed lexicographic analysis, only collocations present in a dictionary will provide additional co-

compositional characterizations that could reveal direct semantic relations and benefit the characterization of the entry" [14].
Smith [16] examined collocations to detect events related to place and date information in unstructured text.
In this paper we describe two statistical methods for extracting collocations from text corpora written in Modern Greek. The first one is the *mean and variance* method which calculates "offsets" (distances) between words in a corpus and looks for patterns of distances with low spread. The second method is based on the X^2 *test*. In section 3 we focus on the main ideas behind the two methods. Some previous work in the field is also covered. Then, in section 4, the two methods used in this work are described. A short description of the data used for testing and some experimental results are given in section 5. We conclude with a discussion for further work.

3 How to Extract Collocations Using Statistical Methods

The traditional approach to collocations has been the lexicographic one. According to Benson and Morton [1] collocates, the "participants" in a collocation, could not be handled separately. Therefore the task of extracting the appropriate collocates is not predictable, in general, and collocations must be extracted, manually, and listed in dictionaries.

In recent years, statistical approaches have been applied to the study of natural languages and the extraction of collocations. It was partially influenced by the availability of large corpora in machine-readable form. Choueka [3] tried to automatically extract collocations from text, using *N-grams* from 2 to 6 words.

A simple method for finding collocations in a corpus is the *frequency of occurrence*. If two or more words often appear together, we have an evidence for collocation. Unfortunately, the selection of the most frequently occuring *N-grams* does not always lead to very interesting results. For example, if we look for bigrams in a corpus the resulting list will consist of phrases such as: *of the*, *in the*, *to the*, etc. To overcome this problem Justeson and Katz [9] proposed a heuristic that improves the previous results. They pass the candidate phrases through a part-of-speech filter and select only those *N-grams* that are likely to be phrases. Some patterns used for collocation filtering (in this heuristic) are *AN*, *NN*, *AAN* and *ANN*, where *A* stands for adjective, *N* for noun. Although the heuristic is very simple the authors reported significant improvement in the results.

The frequency of occurence-based method works well for noun phrases. However, many collocations involve words in other more flexible relationships. The mean and variance method [15] overcomes this problem by calculating the distance between two collocates and finding the spread of the distribution. Namely, the method computes the mean and variance of the offset (signed distance) between the two words in the corpus. This method makes sense intuitively. If the spread of the distribution is low we have a narrow peaked distribution of offsets and this is an evidence of collocation. On the other hand, if the variance is high the offsets are randomly distributed, i.e., there is no evidence of collocation.

"Mutual information" is a measure for extracting collocations [4]. The term "mutual information" originates from information theory. The term "information" has the restricted meaning of an event, which occurs in inverse proportion to its probability and is often defined as holding between random variables. In the work of Church and Hanks the type of "*mutual information*" is defined as holding between the values of random variables. It is roughly a measure of how much one word "tell us" about the other.

We will describe the main ideas behind the two statistical methods, the *mean and variance method* and the X^2 *test* (pronounced 'chi-square test'). We will also give an alternative formula for the calculation of X^2 statistic in the case of extracting bigrams in the corpus. The X^2 *test* is a well-defined approach in statistics for assessing whether or not something is a chance event. This is in general one of the classical problems of statistics and it is usually couched in terms of hypothesis testing. In our case, we want to know whether two words "occur" together more often by chance. We formulate a null hypothesis H_0 for a sample of occurrences. The hypothesis states that there is no association between the words beyond chance occurrences. We calculate the probability p that the event would occur if H_0 were true. If p is too low (beneath a predetermined significance level $p<0.005$ or 0.001) we reject the H_0 (or retain H_0 otherwise). To determine these probabilities usually we compute the t statistic:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \quad (1)$$

where $\bar{x}$ is the sample mean, s^2 is the sample variance, N the size of the sample and μ is the mean of the distribution if the null hypothesis were true.

If the *t statistic* is large enough we can reject the null hypothesis. The problem with the *t statistic* is that it assumes normally distributed data which is not true in general for linguistic data. For this reason we choose the X^2 *test*, which does not assume normally distributed data. However, for this statistics, various side effects have been observed with sparse data. Dunning and Ted [5] proposed an alternative testing *the likelihood ratios* that works better than X^2, when we have sparse data.

4 Methods for Discovering Collocations

4.1 Mean and Variance

The mean is the simple arithmetic average value of the data. If we have n observations $x_1, x_2, \ldots x_n$, then the mean is given by the form:

$$mean = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad (2)$$

The variance of the n observations $x_1, x_2, \ldots x_n$ is the average squared deviation of these observations about their mean:

$$Variance = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} \quad (3)$$

The standard deviation s is the square root of the variance.

$$s = \sqrt{variance} \quad (4)$$

Let see an example from the Greek Language. Consider the verb *κτύπησε (knocked)* and one of its most frequent arguments, *πόρτα (door)*. Here we have some sentences with these two words:

a) Κτύπησε την πόρτα του (He knocked his door)

b) Κτύπησε δυνατά την πόρτα του (He knocked his door loudly)

c) Κτύπησε τη σιδερένια πόρτα του (He knocked his iron door)

d) Κτύπησε τη σιδερένια και βαριά πόρτα του (He knocked his heavy iron door.)

The words that appear between "*Κτύπησε"* (knocked) and "*πόρτα" (door)* are not fixed, so the distance between the two words varies from sentence to sentence. Counting the frequency of occurrence of "*Κτύπησε"* and "*πόρτα"* at a fixed distance would not work here. In order to have a measure of relationship between "*Κτύπησε"* and "*πόρτα"* we can calculate the *mean and variance* of the offsets (signed distances). For the above sentences we compute the mean offset between "*Κτύπησε"* and "*πόρτα"* according to equation (2):

Mean = ¼ (1 + 2 + 2 + 4) = 2.25

If there was an occurrence of the word "*πόρτα"* before the word "*Κτύπησε" we would* entered it as a negative number. The variance of offsets estimates how much of the individual offset deviate from the mean value. It expresses the spread of the distance between the two words.

Using equation (3) we compute the variance as follows:

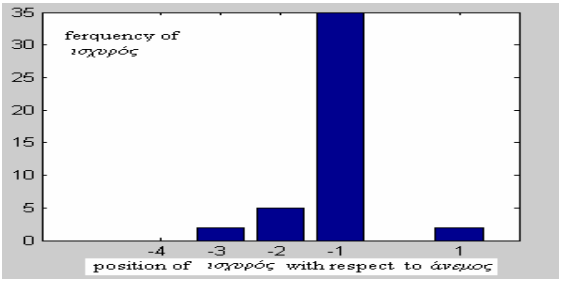
Variance = 1/3 ($(1-2.25)^2 + (2-2.25)^2 + (2-2.25)^2 + (4-2.25)^2$) = 1.58

and the standard deviation is $(1.58)^{1/2} = 1.26$.

Mean and deviation can help us to find collocations by looking for pairs with low deviations. The lower the deviation of the distances between two words the stronger the indication is that these words form a collocation. A low value of deviation means that the two words occur usually at about the same distance.We can explain better the results if we consider the distribution of one word with respect to the other. If there is

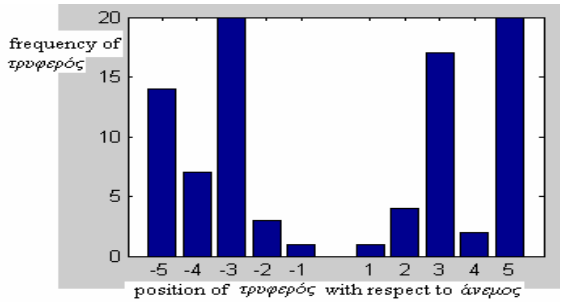
a narrow, peaked spread distribution then this is an indication of a syntactic relation between them. Let explain this case by an example. We count, in our corpus, the occurrences of the Greek words (ισχυρός, άνεμος), calculate the distances of the word "*ισχυρός" (strong)* with respect the word "*άνεμος" (wind)* and take the distribution of Fig. 1:

Fig. 1. Distribution of the distances the word '"*ισχυρός" (strong)* with respect the word "*άνεμος" (wind)*



This is a good indicator of a syntactic relation between the two words as far as we have a peaked distribution with low spread. Whereas, the distribution in fig 2 for the pair of words (*τρυφερός-soft, άνεμος-wind*) does not indicate the same.

Fig. 2. Distribution of the distances of the word *τρυφερός-soft, άνεμος-wind*



So far, we have not discussed anything about extreme values of mean and variance. Unfortunately, this simple method can be a failurein the case of very high frequencies or very low variances. However, we can avoid extreme values taking into account only normal values in comparison with the size of the corpus. We describe next the second technique used here, the X^2 *test*.

4.2 Pearson's chi-square test

In 1900, Karl Pearson developed a statistic that compares all the observed and expected numbers when the possible outcomes are divided into mutually exclusive categories. The form in eq. 5 gives the chi-square statistic:

$$X^2 = \sum \frac{(observed - expected)^2}{expected} \quad (5)$$

Where the Greek letter **Σ** stands for summation and is calculated over the categories of possible outcomes.

The *observed* and *expected* values can be explained in the context of hypothesis testing. If we have data that are divided into mutual exclusive categories and form a null hypothesis about that data, then the expected value is the value of each category if the null hypothesis is true. The *observed* value is the value for each category that we observe from the sample data.

The chi-square test is a remarkably versatile way of gauging the significance of how closely the data agree with the detailed implications of a null hypothesis.

To clarify things we can use an example. Suppose we have a sample of linguistic data and we are interested in extracting collocations of bigrams. Defining a collocational window we count the frequency of occurrences for the pair (*ισχυρός-strong, άνδρας-man*). There are 10 occurrences of *ισχυρός άνδρας* in the corpus, 1000 bigrams where the second word is *άνδρας* but the first word is not *ισχυρός*, 500 bigrams with the first word *ισχυρός* and a second word different from *άνδρας*, and 1,500,000 bigrams that contain neither word in the appropriate position. It would be useful to use the contingency table I in which the data are classified.

Table 1. Contigency table of frequencies fort he word pair (*ισχυρός, άνδρας*)

	w_1 = *ισχυρός strong*	w_1 *ισχυρός*
w_2 = *άνδρας man*	10 (*ισχυρός άνδρας*) strong man	1000 e.g. (σεμνός άνδρας decent man
$w_2 \neq$ *άνδρας*	500 e.g. (*ισχυρός άνεμος*) strong wind	1,500,000 e.g. (ασθενής ήχος) weak sound

Moreover, using maximum likelihood estimates we can compute the probabilities of "*ισχυρός"* and "*άνδρας"* as follows. In our corpus, "*ισχυρός"* occurs 510 times, "*άνδρας"* 1010 times, and there are 1,501,510 tokens overall.

P(*ισχυρός*) = 510/1,501,510.

P(*άνδρας*) = 1010/1,501,510.

The null hypothesis is that occurrences of the words "*ισχυρός"* and "άνδρας" are independent.

H_0 : *P*(*ισχυρός*, *άνδρας*) = *P*(*ισχυρός*) x *P*(*άνδρας*)
=(510/1,501,510)x(1010/1,501,510) ≈ 1.013×10^{-5}

Then we calculate the X^2 value using equation (4). Looking up the X^2 distribution from tables or by using appropriate statistical software, we find a critical value for a significance level *a* (usually *a=0.05*) and for one degree of freedom (the statistic has one degree of freedom for a 2x2 contingency table). If the calculated value is greater than the critical value we can reject the null hypothesis that "*ισχυρός"* and "*άνδρας"* occur independently. So, for a large calculated X^2 value we have a strong evidence for the pair of words. It is a good candidate for a collocation.

We give now a simpler form that we use for calculation of X^2 values. For a contingency 2-by-2 table, equation (5) can take the following form:

$$X^2 = \frac{N(a_{11}a_{22}-a_{12}a_{21})^2}{(a_{11}+a_{12})(a_{11}+a_{21})(a_{12}+a_{22})(a_{21}+a_{22})} \quad \textbf{(6)}$$

Where a_{ij} are the entries of the contingency 2-by-2 table *A* and *N* the total sum of these entries. Implementing this formula in C/C++ programming we must take care of avoiding memory overflow as we divide in (5) the numerator by the denominator. So, to overcome this problematic situation especially when the corpus is very large and the frequencies very small, we separate the total division into subparts by factorizing the formula.

5 Experimental results

Several files of Greek language texts available to us, where a preliminary part-*of-speech* tagging process had been done over them were combined to make a 8,967,432 total linguistic Corpus. The result is considered a very large corpus and it will be very useful resource for future works. Unfortunately, for all of these word occurrences the lemma is provided in our corpus only in the 8,977,083 (or in 30.39%) out of the cases.

We are interesting only for the lemmas where the part-of-speech tag is Noun (No), Verb (Vb), Adjective (Aj) and Adverb (Ad).

These are distributed as follows:

Table 2. Distribution of the lemmas in our corpus

Nouns (No)	6,739,006
Verbs (Vb)	0
Adjectives (Aj)	2,228,426
Adverbs (Ad)	0
TOTAL	8,967,432

Note that lemmas for Verbs and Adverbs are not provided. The remaining *8,977,083-8,967,432=9,651* lemmas belong to a category tagged as RgFwGr. Probably, they concern foreign words that are used identical in Greek Language.

The 10 most frequent nouns and adjectives in the Corpus are shown below:

Table 3. 10 most frequent nouns and adjectives in the Greek Corpus

	Noun Frequency			Adjective Frequency	
1	"ελλάδα"			"πολιτική"	25892
2	60318			"μεγάλη"	15965
3	"νόμος"			"περισσότερος"	15147
4	33680			"ελληνική"	13508
5	"κανόνας"			"νέος"	12744
6	31349			"εθνική"	11360
7	"θέση"			"μεγάλος"	10929
8	31011			"όλη"	10680
9	"διεθνός-διεθνώς"			"οικονομική-οικονομικής"	10664
10	30369			"ελληνικός"	9057
	"πρόβλημα"				
	27835				
	"κυβέρνηση"				
	26095				
	"χρόνι-χρόνια-χρόνιο"				
	25580				
	"πρόεδρος"				
	25302				
	"θέμα"				
	25297				

The quotes are used here to denote that these word forms represent lemmas as exactly they appear in the corpus files.

5.1 Analysis of Variance

The only combination of bigrams we could try is that of pairs (Adjective, Noun) as the files don't provide the other part-of-speech.

We calculate from the corpus the distances and the standard deviation of these distances for all the combinations of bigrams (Adjective, Noun), defining a collocational window of 10 words (including punctuation marks). Remember that for a positive distance d ($-10 <= d <= 10$) we denote that the noun is found in a distance of d words longer on the right hand side of the adjective. A negative distance denotes the opposite.

In the two tables 4 and 5 below are shown the standard deviation and the distances for the 2 lowest and the 2 highest standard deviation bigrams of our calculations. In each entry they appear sequentially: The words of the bigram, the frequency of the bigram in the corpus, the mean value of the distances, the standard deviation of the distances. In the figures 3 and 4 we plot the distributions of distances for the first lowest standard deviation bigram and the first highest standard deviation bigram. While in tables 6 and 7 they are shown the 10 lowest standard deviation bigrams and 10 highest standard deviation bigrams respectively.

Table 4. The two lowest standard deviations bigrams

(adj,Noun),	frequency,	mean,	std
("χρονικό","διάστημα")	1983	0.9561	0,7654
("κεντρική","σημασία")	13	1.2308	0,8321

Table 5. The two highest standard deviations bigrams

(adj,Noun),	frequency,	mean,	std
("εξωτερική-εξωτερικό-εξωτερικός", "τρόπος")	17	1.2353	8,2956
("εσωτερική""γιώργος")	12	0.9167	8,6072

Fig. 3. Distribution of distances for the lowest standard deviation bigram (“*χρονικό”,”διάστημα”*)

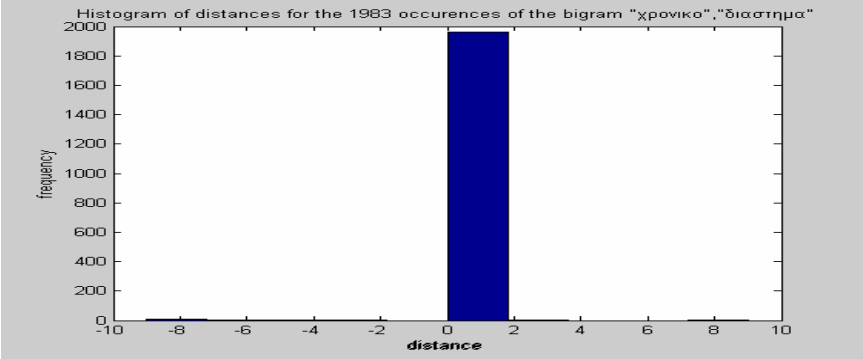


Fig. 4. Distribution of distances for the highest standard deviation bigram ("εξωτερική-εξωτερικό-εξωτερικός", "τρόπος")

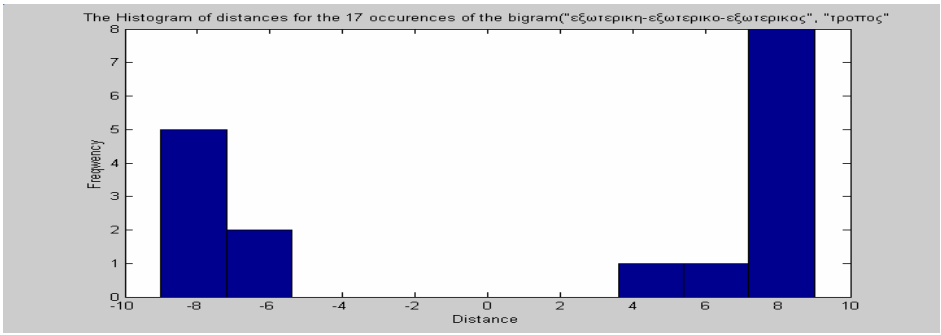


Table 6. The 10 lowest standard deviation bigrams in the corpus

Lemmaadj	lemmanou	stdv
"χρονικό"	"διάστημα"	0,7654
"κεντρική"	"σημασία"	0,8321
"ειδικός"	"απάντηση"	1,1875
"μεγάλος"	"βαθμός"	1,1932
"περασμένος"	"κανόνας"	1,3007
"αμερικανική-αμερικανικής"	"κανόνας"	1,3817
"κυριακής"	"ελλάδα"	1,3901
"ανά"	"κόσμος"	1,4151
"οικονομικό"	"παιχνίδι"	1,4434
"εργαζομένα-εργαζομένη-εργαζομένης"	"διεθνός-διεθνώς"	1,4546

Table 7. The 10 highest standard deviation bigrams in the corpus

Lemmaadj	lemmanou	stdv
"ζήτημα"	"πληροφορία-πληροφορίη"	7,854
"χθεσινής"	"συμμετοχή"	7,866
"ανά"	"ιστορία"	7,8671
"μήνα-μήνη"	"χρήση"	7,8758
"ενδεχόμενος"	"θέμα-θέμας"	7,8988
"χθεσινής"	"διεθνός-διεθνώς"	7,9036
"εθνικός"	"μείωση"	7,9174
"συγκεκριμένο"	"ιστορία"	7,9601
"κοινωνική-κοινωνικής"	"λαός"	7,9663
"λίγη"	"διεθνός-διεθνώς"	7,9935

Interpretation: For a bigram with a low standard deviation of the distances between the words and a half-sided high peak value distribution is a strong indication for these words to form a collocation. The narrower the shape and the higher the peak value of the distribution are the stronger the indication is for these words to form a collocation.

5.2 Analysis of *X-square* test

The *X-square* test is more flexible from that of variance, which it can be accidental in the cases of extremely high frequencies. The X^2 statistic makes a hypothesis (the null hypothesis) of statistical independence for the two words of a bigram. That is, the null hypothesis supposes that the two words occur independently of each other within the corpus. Calculating the X^2 statistic we can reject the null hypothesis if it exceeds a critical value as defined from the X distribution. For example, if we look up a statistical table (or the returned value from a statistical software), we find that at a probability level of a=0.05 (that means 95% sure) and one degree of freedom (the statistic has one degree of freedom for bigrams, namely, for a 2-by-2 contingency table) the critical value to reject the null hypothesis is X^2= 3.841.

Of course we could try for a higher significance level of 99% (a=0.001) or more, but this would increase the corresponding critical value. The essence of the test is to examine the X^2 calculation and to decide for the dependence of the two words. The higher the value of *X-square* statistic is the stronger the indication will be for these words to be dependent.

Experimental results. Our Corpus consists of 29,539,802 words. Given this number and a collocational window of 10 words around a target adjective we can calculate the total number of bigrams (*adjective,noun*). This can be calculated by the form:

Total number of bigrams=(29539802-9)*9+36

For each one of these bigrams we scan the corpus and calculate the a_{ij} entries of the 2-by-2 contigency table to evaluate eventually the X^2 score.

Tables 8 and 9 show the 10 highest X^2 score and the 10 lowest X^2 score bigrams respectively. In second table the X^2 score is not exactly zero but approximates zero.

Table 8. The 10 highest *X*–square score bigrams in the corpus

Adjective	Noun	X2score	a11	a12	a21	a22
"κοινωνικής"	"διάλογος"	3,4057	59	117373	41737	265699004
"κοινωνικής"	"μείωση"	3,3964	10	112994	41786	265703383
"διαφορετικός"	"μέλη"	3,3488	11	116863	43135	265698164
"συγκεκριμένος"	"σημασία"	3,3426	11	111553	45637	265700972
"χρονικό"	"δημόσια"	3,3325	9	112041	41553	265704570
"προοπτική-προοπτικής"	"μείωση"	3,2941	11	112993	45169	265700000
"ίδιο"	"παρουσία"	3,1651	11	112471	44161	265701530
"διαφορετικός"	"συμφωνία"	3,1563	11	115063	43135	265699964
"σημερινή"	"συμφωνία"	3,1501	10	115064	42776	265700323
"κυπριακός"	"σημασία"	3,1498	12	111552	47454	265699155

Table 9. The 10 lowest *X*–square score bigrams in the corpus

Adjective	Noun	X2score	a11	a12	a21	a22
"σημερινή"	"στιγμή"	0	16	299972	42770	265515415
"δηλώσης"	"πρόγραμμα"	0	15	280533	121305	265456320
"ουσιαστικά"	"σημείο"	0	18	190638	70182	265597335
"εξωτερική"	"ευρωπαϊκή"	0	251	227377	160849	265469696
"έτοιμος"	"λόγος"	0	21	367683	50235	265440234
"εσωτερική"	"ενδιαφέρον"	0	5	149125	66505	265642538
"μεγάλος"	"άτομο"	0	50	133114	196672	265528337
"ίδιο"	"παιχνίδι"	0	19	211625	44153	265602376
"αργότερα"	"τουρκία"	0	41	299191	108085	265450856
"οικονομικός"	"προσπάθεια"	0	18	239634	55746	265562775

6 Discussion and further work

This work presents 2 methods of fully automatic finding collocational words for the Greek language. The "mean and variance" method and the X^2 testing. In the last case, we have demonstrated with the experimental results that it is possible to work fine with large corpora of the Greek Language. In respect of the use of various NLP tools the methods are self-sufficient.

Although in calculating significance we could use many other well suited for this purpose statistical methods, like mutual information (*MI*), log likelihood (*LL*) ratio test, *t-test* etc., these methods were rejected in favor of the chi-square statistics. The reason is that these tests have the important flaw that they assume a parametric distribution of the data. This is manifestly unsuitable when counting frequencies of bigrams. Moreover, *MI* compares the joint probability $p(w_1,w_2)$ that two words occur together with the independent probabilities $p(w_1)$, $p(w_2)$ that the two words occur at all in the data, $MI(w_1, w_2) = log_2(p(w_1,w_2) / p(w_1)* p(w_2))$ and this does not give realistic figures for very low frequencies. If for example a relatively unfrequent word has a frequency of 1 in a certain combination, the resulting very high value of *MI* may lead to a decision for a strong link between the words although cooccurrence might be simply by chance.

The X^2 testing is the most commonly used test of statistical significance in computational linguistics and can be used here in many different contexts.

The following are directions for a future word.

The system can incorporate lexical knowledge to assist in finding collocations and improving the results. Such knowledge may be come from the use of lexical thesaurus, synonymy, hypernymy and part of speech tagging available for the Greek language. Pearsen [13] has worked in a similar way using WordNet Lexicon [12] for the English language. Using these statistical methods we believe that we might just

get a very good representation of the prepositional knowledge. By combining statistical methods in a conceptual graph knowledge representation framework, we could collect valuable information finding semantically related words and thus obtain richer knowledge bases. Computer assessment of knowledge structure combining statistical methods seems to be an interesting and important next step.

7 References

1. Benson & Morton 1989. The structure of the collocational dictionary. In International Journal of Lexicography 2:1-14.

2. Caroll J., Minnen G., Pearse D., Canning Y., Delvin S. and Tait J. (1999). Simplifying text for language-impaired readers. In Preceedings of the 9th Conference of the European Chapter of the ACL (EACL '99), Bergen, Norway, June.

3. Choueka, Y.; Klein, T.; and Neuwitz, E. (1983). "Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus." Journal for Literary and Linguistic Computing, 4, 34-38. In Information Theory, 36(2), 372-380. Fano, R. (1961). Transmission of Information: A Statistical Theory of Information. MIT Press. Flexner, S., ed. (1987). The Random House.

4. Church, K., and Hanks, P. (1989). "Word association norms, mutual information, and lexicography." In Proceedings, 27th Meeting of the ACL, 76--83. Also in Computational Linguistics, 16(1). algorithm." IEEE Transactions on Information Theory, IT-26(1), 15-25. HaUiday, M. A. K., and Hasan, R. (1976). Cohesion in English. Longman.

5. Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, Volume 19, number 1, pp61-74.

6. Firth J. R. (1957). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp 1-32. Oxford: Philological society. Reprinted in F. R. Palmer(ed), Selected papers of J. R. Firth 1952-1959, London: Longman, 1968.

7. Gitsaki C., Daigaku N. and Taylor R. (2000). English collocations and their place in the EFL,classroom available at: http//www.hum.nagoya-cu.ac.jp/~taylor/publications/collocations.html.

8. Howarth P. and Nesi H. (1996). The teaching of collocations in EAP. Technical report University of Leeds, June.

9. Juteson S. and Katz S. (1995b). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Languagr Engineering 1:9-27.

10. Lin D. (1998). Extracting collocations from text corpora. In First Workshop on Computational Terminology, Montreal, Canada, Augaust.

11. Manning C. and Schutze H.(1999). Foundations of Statistical Natural Language Processing (Fifth Printing 2002). The MIT Press.

12. Miller G., Beckwith R., Fellbaum C., Gross D. and Miller K. (1993). Introduction to WordNet: An On-line Lexical Database. Five Papers on WordNet Princeton University.

13. Pearce D. (2001). Synonymy in Collocation Extraction. . In WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop). pages 41-46. June. 2001. Carnegie Mellon University, Pittsburgh.

14. Richardson, S. D. (1997). Determining similarity and inferring relations in a lexical knowledge base [Diss], New York, NY: The City University of New York.

15. Smandja F. (1993). Retrieving collocations from text: Xtract. Computational Linguistics, 19(1):143-177, March.

16. Smith A. David (2002). Searching across language, time, and space: Detecting events with date and place information in unstructured text July 2002 In Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries