Problems of Statistics in the Greek Linguistic Studies: A Computational Approach

Abdelmonem Ahmed Zaki[*1]
Faculty of Arts, Ain Shams University
abdelmoneam.ahmed@art.asu.edu.eg

Abstract: ***This paper aims to explain why we would use statistical methods for Ancient Greek linguistics? The statistics are typically using a large machine-readable corpus, in order to discover general principles of linguistic behavior, genre difference, etc. The paper also sets out to prove some hypotheses, or identify some linguistic phenomena, such as morphological, syntactic, and semantic phenomena.***

The proposed paper will consist of the following points:

- ***What kinds of linguistic data can they handle?***
- ***What are the advantages and disadvantages of statistical linguistics?***
- ***What is the nature of the assumptions they require of the analyst?***
- ***What is the strategy for studying of linguistic phenomena?***

Keywords: ***Statistical Linguistics, Linguistic Phenomenon, Greek Linguistics, Computational Approach. Morphological Classification, Semantic Classification.***

1 INTRODUCTION

In 1959, N. Chomsky said that statistical approaches will always suffer from lack of data, and that language should be analyzed at a deeper level [1]. In 2007, he said also that it was commonly assumed that statistical analysis of vast corpora should reveal everything there is to learn about language and its acquisition [2].

So Chomsky believes that the statistical analysis is not enough to study the language, and here begins the issue, statistics suitable for studying all linguistic phenomena? Is it possible to rely on statistical results in the formation of linguistic base? And why do linguists need statistical analysis?

A. Problem of the Study:

Many linguistic studies are based on statistical analysis, but, do the statistics fit for the study of all linguistic levels such as: phonology, morphological, syntactic, and semantic? And, is it possible to rely on statistical results in the formation of linguistic base?

B. Aims of the study:

- To present a methodology for statistical analysis of ancient Greek texts.
- Propose a computational system that includes a data base of Greek vocabulary, to help researchers in understanding the different Greek linguistic phenomena through statistics.

C. Axes of the study:

The researcher divides the study into two levels:

- Level I: Statistical analysis of the Greek language problems.
- Level II: foundations that must be followed to create a computer system capable of statistical analysis of the Greek language.

D. Resources of the study:

The researcher relied on the Thesaurus Graecae Linguae (TLG), which is a comprehensive library of the most ancient Greek texts.

E. Computational tools:

- Search programs in classical texts, like, Musaios, and Diogenes.
- The researcher used the Perseus site to study the classical texts [3]
- The researcher also used the analyzes N-gram [i] through the site: Http://guidetodatamining.com/ngramAnalyzer

2 WHAT IS STATISTICAL LINGUISTICS?

Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances [4]. Thus, statistics is an indicator of how 'correct' your results are, if you have based the calculations on appropriate assumptions and interpreted the results correctly. Finally, statistics can be used in two ways: to describe a data set, or to draw inferences

outside of the data set (descriptive and inferential statistics, respectively). The conditions for describing or drawing inferences are obviously not the same, and this means that it is important to define what is being studied, and how the conditions for a given test are met in the data set [5]. The notion of data type is crucial to all branches of statistics. Because all statistical tests make assumptions about types of data (they are quite picky). In corpus linguistics, we are almost always dealing with nominal data. [6] Language is a collection of statistical distributions: Weights for rules (phonetic, syntactic, etc) change when learning, a long time, between communities. The statistical work shows us the reality of language in specific stage or in two periods. Statistical methods are relevant to language acquisition, change, variation, generation and comprehension.

3 WHAT IS THE COMPUTATIONAL APPROACH?

The Computational Approach is the ability of a computer-system to be ***"Self-Aware"*** in some way. More specifically, we mean the property of a computational or formal system to be able to access and internalize some of its own properties.

4 PROBLEMS OF STATISTICAL ANALYSIS IN LINGUISTIC STUDIES:

A. Text Classification.
B. Morphological Statistical Classification.
C. *Semantic Statistical Classification*

A. Text Classification.

There is no doubt that statistical analysis is of crucial importance to verify the linguistic phenomenon. But could we make statistical analysis in the same manner to all of the TLG authors? For example, if we search the frequency of the word "disease" "ἡ νόσος" it is obvious that the word is more commonly used at medical scientists, unlike poets or historians. So in the following statistics in TABLE 1, we'll limit the highest proportion of recurrence at three authors only:

TABLE 1
THE FREQUENCY OF THE WORD "ἡ ΝΟΣΟΣ "IN THE GREEK TEXTS

Authors	Vocabulary	Frequency	Corpus	Percentage	Date
Aretaeus	60598	395	Corpus MedicorumGraecorum, De causis et signis acutorum morborum	0.65%	II b.c
Hippocrates	78666	286	Corpus Hippocraticum, De Diaeta In Morbis Acutis	0.36%	IV/V b.c
Josephus	305870	106	Antiquitates Judaicae	0.034%	I a.d

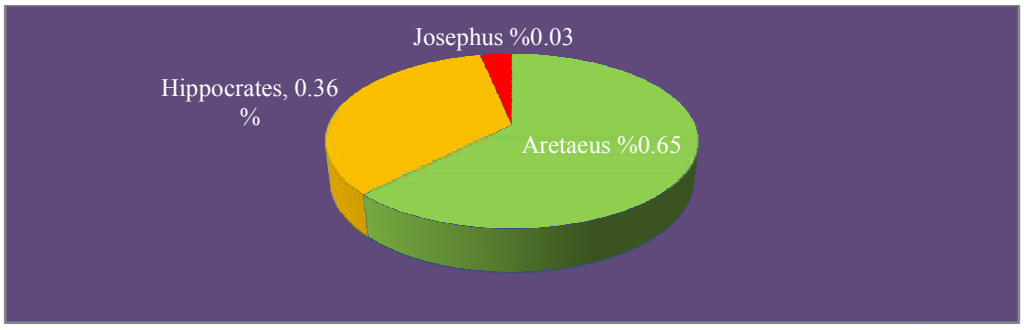

Figure 1: The frequency of the word "ἡ νόσος"

We conclude from the Figure 1 that Aretaeus is the most widely used of the word **"ἡ νόσος".** So Josephus should be excluded from the comparison, as his writings were not about medicine, but the comparison between Aretaeus and Hippocrates is true. Anyway the comparison should be between two authors or more in the same topic.

So when we search again the frequency of the word "night" **"ἡ νύξ"** at the two Greek Historians, Diodorus Siculus and Cassius Dio (see TABLE 2)**.**

TABLE 2

THE FREQUENCY OF THE WORD"ἡ νύξ" BETWEEN DIODORUS SICULUS AND CASSIUS DIO.

Authors	vocabulary	frequency	Corpus	Percentage	Date
Diodorus Siculus	271501	255	Bibliotheca Historica	0.094%	I b.c
Cassius Dio	399409	165	Historiae Romanae	0.041%	II a.d

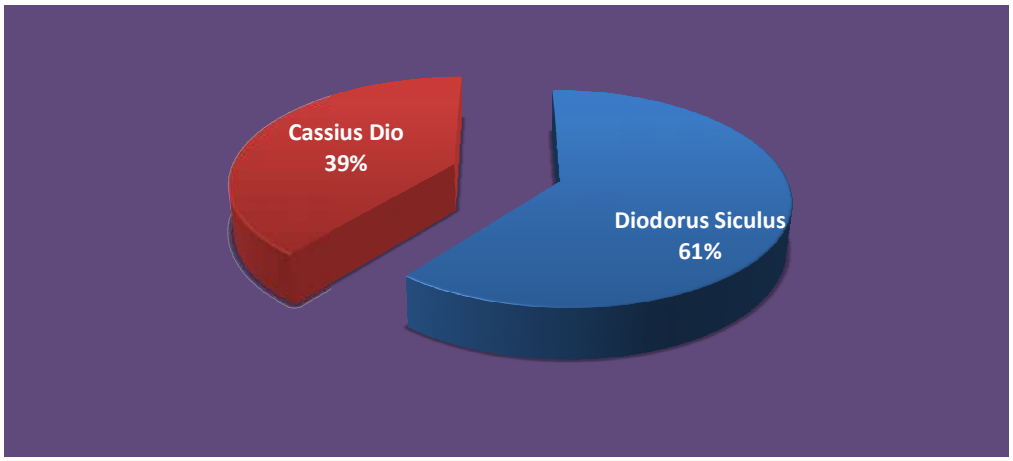

Figure 2: The frequency of the word "ἡ νύξ"

We conclude from the Figure 2 that the comparison between Diodorus Siculus and Cassius Dio is true. But we should organize the results historically from the past to the present, not form the most widely used.

B. Morphological Statistical Classification.

- Multi-Function Morphemes:

The morphological analysis of the texts is important in computational processing in general. The morphological statistical analysis should be done for vocabulary based on distinguishing nouns, adjectives, verbs, prepositions, particles ... etc., The researchers faced some morphological difficulties when counting morphemes, due to Multi-Function Morphemes. The Multi-Function Morphemes are big problems for the Arab students when studying ancient Greek language. Because the Greek Language has 15 multi-functional morphemes : {ει}, {θεντων}, {ων}, {σαι}, {ῃ}, {οι}, {ος}, {ουσι}, {ετε}, {α}, {ον}, {οιν}, {ε}, {ου}, {σθον}, and five of these morphemes are responsible for 60% Morphological Mistakes of Arab Students: {ει}, {θεντων}, {ων}, {σαι}, {ῃ}.[7]

So the Multi-Function Morphemes pose also a problem for the current programs that search in TLG like Musaios (see Figure 3), TLG, Diogenes (see Figure 4), it's difficult to distinguish suffixes from nouns or verbs or particles …etc, it searches only in word, without consideration if it is a noun or verb….etc, and gets the frequency without any linguistic classification.

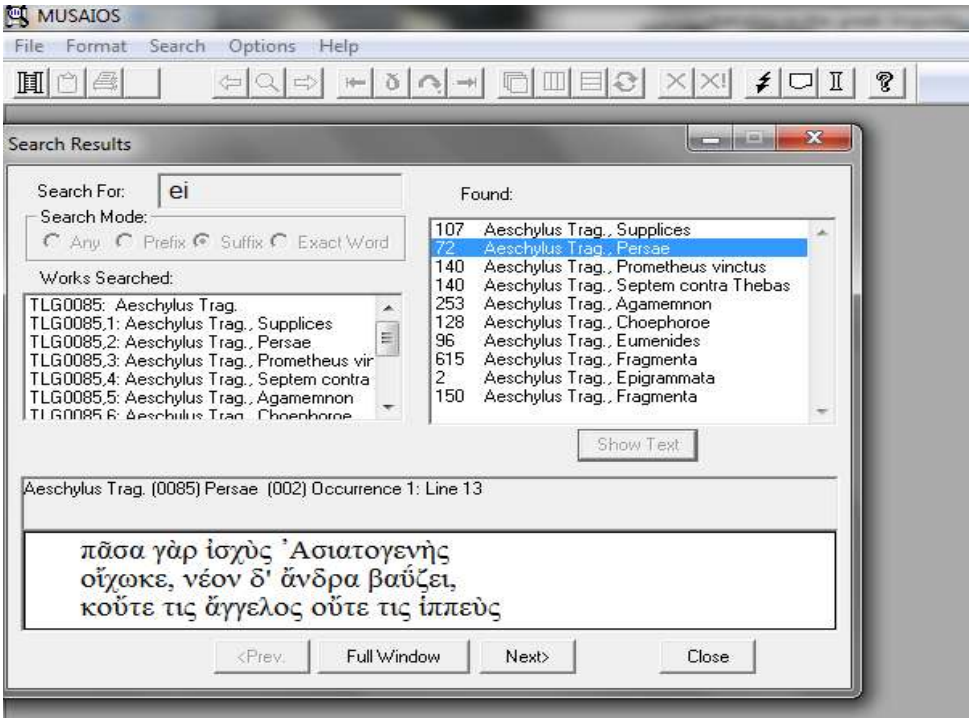

Figure 3: Searching the morpheme {ει} by Musaios

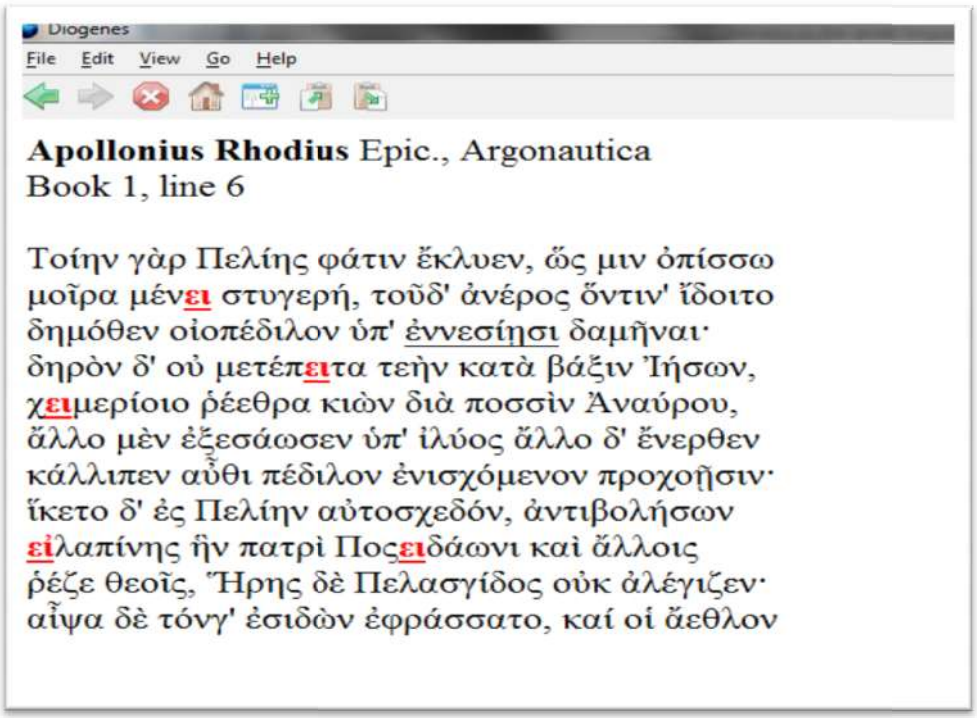

Figure 4: Searching the morpheme {ει} by Diogenes

For example the morpheme {ει} is a multi-function morpheme used as:

- A suffix for Noun 3rd. Decl. Sing. Dat. Like πόλει.
- A suffix for Verb, Indic. Act. Pres. 3rd.pers. sing. Like λύει
- A suffix for Verb, Indic. Act. Fut. 3rd.pers. sing. Like λύσει
- A suffix for Verb, Indic. Mid &Pass. Pres. 2nd.pers. sing. Like λύει
- A suffix for Verb, Indic. Mid. Fut. 2nd .pers. sing. Like λύσει.
- A suffix for Verb, Indic. Pass. Fut. 2nd .pers. sing. Like λυθήσει.
- A suffix for Verb, Indic. Act. Perf. 3rd Pers. Sing. Like ἐλελύκει.
- Verb to be, Indic. Act. Pres. 2nd Pers. Sing. Like εἶ.
- The Adverb ἀεί ends with the diphthong ει which can't be distinguished by programs like Musaios.
- The conditional particle εἰ which was the same like the morpheme {ει}.

Practical Example:

The researcher performs a statistical analysis of the morpheme {ει} through the Persians of Aeschylus by using Musaios, and the results were as follows:

- The morpheme {ει} was used 72 times in the Persians of Aeschylus (see Figure 5), without any morphological classification. So the researcher suggests to create a computer-system which includes a Data-Base of Greek vocabulary that can classify the Greek Texts by morphemes and be able to search and distinguish between texts.
- So the results can be classified as in TABLE 3:

And the computer system should summarize the results as follows:

- Verbs: 43 (see TABLE 3)
- Nouns: 12 (see TABLE 4)
- Adjectives: 2 (see TABLE 4)
- Particles And Adverbs: 12 (see TABLE 5)

- *Morpho-phonoligical Changes:*

Most inflection in Greek consists of adding an ending to a fixed stem. Greek shows great freedom in forming compound verbs by the addition of prepositional prefixes. From the stem γραφ- is derived παρα-γραφ-ω, κατα-γραφ-ω, ὑπο-γραφ-ω, etc. It would be uneconomical to include παραγραφ-, καταγραφ-, and ὑπογραφ- in the dictionary since all are formed by the addition of common prefixes to the single verb stem γραφ-. A difficulty arises, however, from the fact that the prefixes are often assimilated phonetically to the following letter. The prefix *συν-* ' together' appears as συν- before vowels and dental consonants, as συμ- before labial consonants, as συγ- before guttural consonants, as συλ- before λ, and as συ(σ) before σ. The prefix μετα- appears as μετα- before consonants, μεθ- before vowels with aspiration and μετ- before vowels without aspiration. The program must recognize the assimilated forms of each prefix and must verify that the letter following the prefix could in fact have caused the suspected assimilation. In some cases a single verb is compounded with as many as three prefixes, each of which may appear in an assimilated form. The form συγκαθίστημι must be analyzed as ο συν + κατα+ ἵστημι , συνεπανίσημι as συν + ἐπι+ ἀνα+ἵστημι. Further complication is caused by the fact that verbal augments come before the stem but after the prefixes. The imperfect of συμ-βαίν-ω is συν-έ-βαιν-ον. Thus, if the word cannot be analyzed directly into a stem and an ending, the program must attempt to remove prepositional prefixes from the beginning of the word. If a hypothetical prefix can be removed, the program proceeds to analyze the remainder of the word. If this analysis is successful the prefix is reunited with the word in the final analysis. In some cases the program makes more than one hypothetical division between prefix and stem. The verb ἀναλύω would generate three hypothetical divisions: ἀνα-ἀλύω + ἀνα ἁλύω + and finally ἀνα-λύω.[8]

TABLE 3
VERBS IN PERSIANS OF AESCHYLUS

Serial	Verb	Origin	Verse
1	Βαύζει	Βαύζω	13
2	Πέμπει	Πέμπω	54
3	ἐλαύνει	ἐλαύνω	75
4	ἐπάγει	ἐπάγω	85
5	Παράγει	Παράγω	111
6	Πρέπει	Πρέπω	239, 247
7	κἀπιδεσπόζει (και ἐπιδεσπόζει)	ἐπιδεσπόζω	241
8	Φέρει	Φέρω	248
9	ἤρκει	ἀρκέω	278
10	ὑπερβάλλει	ὑπερβάλλω	291
11	Βλέπει	Βλέπω	299
12	ἔχει	ἔχω	343, 597, 724
13	Πλήθω	Πλήθει	352
14	προφωνεῖ	Προφωνέω	363
15	ἐπῄει	ἔπειμι	378
16	ἐχώρει	Χωρέω	379
17	Παρεκάλει	Παρεκάλω	380
18	κυρεῖ	Κυρέω	598
19	ἐκφοβεῖ	ἐκφοβέω	606
20	ἀίει	ἀίω	633
21	Κλύει	Κλύω	639
22	ἀνίει	ἀνίημι	649
23	Ποδούχει	Ποδουχέω	656
24	πονεῖ	Πονέω	682
25	ἐστρατηλάτει	στρατηλατέω	717
26	Στένει	Στένω	730
27	ἔρρει	ἔρρω	732
28	κρατεῖ	Κρατέω	738
29	Μνημονεύει	Μνημονεύω	783
30	Πέλει	Πέλω	792
31	συμμαχεῖ	Συμμαχέω	793
32	Κυρήσει	Κυρέω	797
33	Συμβαίνει	Συμβαίνω	802
34	Λείπει	Λείπω	804
35	ἄρδει	ἄρδω	806
36	ἐπαμμένει	ἐπαναμένω	807
37	ὠφελεῖ	ὠφελέω	842
38	Δάκνει	Δάκνω	846
39	ἀμπέχει	ἀμπέχω	848
40	αἰάζει	αἰάζω	922
41	ἀυτεῖ	ἀυτέω	1058

TABLE 4
NOUNS &ADJECTIVES

Serial	Noun	Origin	Verse
1.	ἀλύξει	ἄλυξις	108
2.	πολεῖ	Πόλις	307, 715, 781
3.	Τάχει	Τάχος	342
4.	πενθεῖ	Πένθος	579
5.	Στένει	Στένος	683
6.	Τάρβει	Τάρβος	696
7.	εὐτυχεῖ (adj.)	εὐτυχής	709
8.	Δαρεῖ	Δαρεῖος	713
9.	βραχεῖ (adj.)	*Βραχύς*	713
10.	Θράσει	Θράσος	744
11.	φρονεῖ	Φρόνις	782
12.	Θράσει	Θράσος	831

TABLE 5
PARTICLES & ADVERBS

Serial	Adv./Particles	Verse
1	κεὶ	295
2	ἐκεῖ	319
3	εἰ	357, 369, 631, 790, 791, 800
4	ἐπεί	377, 656, 697, 703

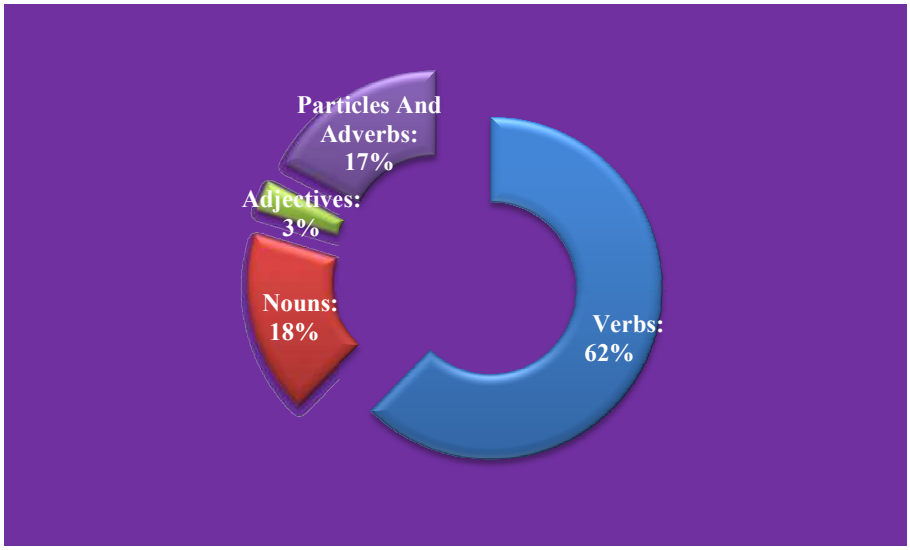

Figure 5: The Morpheme {ει} in the Persians of Aeschylus

C. Statistical Semantic Classification for Greek Texts

What is the content of the text? Or to be more precise: What are the basic elements that should be identified in order to understand the basic meaning of the text? Whether poetry or prose. Every text contains some basic phrases that constitute the framework: So the problem is to identify the central core of the text that carries meaning basics. This is the first critical step, before any attempt at explanation can be provided. If there is no program capable to perform semantic analysis of Greek texts, how it can be automated semantic classification of Greek texts? Can it make a comparison of Greek texts to determine the influence among poets and writers? The current computer programs can't distinguish semantic fields of Greek texts. But we can't measure some semantic features for Greek writers as activity and optimism and certainty, realism and commonality.

The Researcher proposes to create a computer system for the analysis and statistical classification of Greek texts
- The system should identify most Greek words automatically.
- Computer System should contain different combinations of Greek lexicons, and supports a database of Greek vocabulary which shows the stages of semantic development of the Greek vocabulary through the ages and in different contexts.
- The system should support different files with extensions such as: (.txt, doc., .docx., .html, .xml, ... Etc)
- The system should give the users multiple options for counting and classification of linguistic data.
- The system should allow users to compare the results among authors.
- The system should also allow users to choose dictionaries, which can be compared through texts, to give further semantic analysis
- The system should also display the results in statistical tables and charts.

5 CONCLUSION

The Researcher finds through the study that:

- Greek texts can be processed automatically in the light of digital content, which, now, is online.
- Using computer in counting different language phenomena, but there are some problems in morphological processing due to the presence of multiple morphemes function as mentioned above.

- No one can deny the benefits of statistics in general linguistic studies, but there are regulations to be followed when studying the any linguistic phenomenon including the following:
    a) Selection of the type of text (literary / poetic / scientific / .... etc).
    b) The selection of the author (poet / historian / doctor / ..... etc).
    c) Choose period (the same period / two periods/ etc ...).
    d) Choose the means of research and comparison.
    e) Comparing texts in different semantic contexts.

REFERENCES

[1] http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman
[2] N-gram is A mechanism analyzes to determine the frequency of the word Uni gram, or phrase Bi gram & Tri gram or collocations in the corpus, with the presentation of the results sorted by number of iterations, and said ratio to the number of words of the corpus. http://nlpwp.org/book/chap-ngrams.xhtml (accessed 15 November 2014).
[3] L. Padrò, *Statistical Methods for Natural Language Processing*, Columbia University, 2009, p. 7.
[4] N. Chomsky, *Of Minds and Language,* Biolinguistics 1, 2007, p.10. http://www.biolinguistics.eu (accessed 3 November 2014).
[5] M. Davidian & T. A., Louis, *What Is Statistics?,* American Statistical Association, http://www.amstat.org/careers/whatisstatistics.cfm (accessed 14 November 2014)
[6] G. B. Jenset, *Basic Statistics for Corpus Linguistics*, Handout for methods seminar in English linguistics. Fall 2008. https://www.researchgate.net/publication/265965333_Basic_statistics_for_corpus_linguistics (accessed 1 August 2014).
[7]A. Farrag, (2003) *Redesign the Curriculum for Bilingual Greek and Latin for Arab Student, Critical and applied Study, according to the Theory of Error Analysis in the Light of Educational Linguistics*, Egyptian Society of Comparative Literature, Brochures-Comparisons, Brochure (1), 2003, pp.86-94. (in Arabic)
[8] D. W. Packard, *Computer-assisted Morphological Analysis of Ancient Greek,* Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics, Pisa 21/8 to 1/9/ 1973, Vol. 2, pp.345-346, http://ww.aclweb.org/anthology/C73-2026 (accessed 2 December 2014).

Dr. Abdel-Monem Ahmed Zaki

Director of the Computer Unit, Faculty of Arts, Ain Shams University.

He is Lecturer of computational linguistics in the Department of Ancient European Civilization, Faculty of Arts, Ain Shams University. He obtained his MA in Historical Linguistics in 2007, and his PhD in Computational Processing of Ancient Greek Lexicons in 2012. His main areas of interest are concerned with corpus work, morphological and lexical analysis.

Dr. Abdel-Monem Has many scientific works in Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) American Philological Association - USA, (3) Egyptian Society of Greek and Roman Studies, Cairo (4) The Association of Egyptian Papyrologists, Cairo.

مشكلات علم الإحصاء في الدراسات اللغوية اليونانية "مقاربة حاسوبية"

د/ عبدالمنعم أحمد زكي
كلية الآداب -جامعة عين شمس
abdelmoneam.ahmed@art.asu.edu.eg

ملخص

يحاول الباحث من خلال هذه الدراسة إلقاء الضوء على أهمية علم الإحصاء في الدراسات اللغوية، ويجيب عن بعض الاستفسارات مثل: ما هي مشكلات علم الإحصاء في الدراسات اللغوية؟ وما هي الإستراتيجية المتبعة لدراسة ظاهرة لغوية؟ وما هي البيانات التي يمكن معالجتها؟ وكيف يمكن صياغة الفرضيات والنتائج؟ كما

يضع الباحث الأسس لإنشاء نظام حاسوبي يشتمل على قاعدة بيانات للمفردات اليونانية وذلك لإعانة الباحثين في فهم الظواهر اللغوية المختلفة للغة اليونانية القديمة من خلال علم الإحصاء.