Lexical Computational Models:
The Case of the Byzantine Greek Language

Nikolaos E. Myridis
Assistant Professor, Aristotle University of Thessaloniki
University Campus, GR-54124, Thessaloniki, Greece
Tel: 30-2310-996970 E-mail: nmyridis@theo.auth.gr

Abstract

In this article in-depth analysis and processing of dictionaries' corpora is presented, based on e-collections of them. Moreover special interest is given to the development of means and tools for the coupling of lexica to (speech or textual) linguistic corpora. The relative results are (a) the construction of solutions for language processing based on lexicons' prior information and infrastructure, as well as (b) the development of adequate models for fitting and representation of real, lexical or language, data. Applications of language processing relative to (a) metric measurements (b) distributions and (c) spectra are given. Analytical models, developed in order to fit real measurements in lexica, are constructed, with expansion to language processing. As an instance, the e-corpus of the Medieval Greek Language regarding the period 1100AD-1669AD [expanding from Middle Byzantine era to Late Byzantine period and then to Post-Byzantine years], based on the renowned dictionary of Prof.Kriaras, has been constructed. Elements and results of the aforementioned language processing are given. This kind of processing is performed for the first time regarding the Greek language.

Keywords: Language processing, Sub-corpora, Byzantine studies, (Computational) lexicography

1. Introduction

The purpose of this paper is multiple:

(a) to illuminate the Byzantine (Medieval) Greek Language

(b) to show its crucial importance

(c) to prove also its possible beneficial impact on modern languages

(d) to investigate in-depth structural rules of dictionaries

(e) to extract and construct means and models for lexicographical analysis, and

(f) to apply these models in lexicography-based language processing.

The abundant lexicon of the Medieval Greek Language, based on the work of Prof.Kriaras, is used as reference for our endeavor (Kriaras, 1969-). The work presented in this paper relates also, due to the previous purposes, to computational lexicology, computational lexicography and computational linguistics (Hartmann & James, 1998).

The Medieval Greek Language is obviously Byzantine Greek Language. The lexicon of Kriaras refers to the period 1100AD-1669AD expanding from Middle Byzantine era to Late Byzantine period and then to Post-Byzantine years (Evans, 2004).

Greek is a language with a history of many thousands years (Andriotis, 1995). The extremely long period of its life as well as its various phases during this period are remarkable aspects regarding the development and understanding of this language. In this paper we focus on the specific sub-interval of the years between 1100 and 1669 AD. The Greek language of that period is well known as the *Medieval Greek Language* (we substitute by $\overline{m}GrL$), as it is indicated by the title of Kriaras lexicon. We are interested in this specific phase of Greek language, thanks to the renowned dictionary of Prof. Em.Kriaras, an 104-years old man devoted his life to this work, beginning at 1969, since today.

1.1 The source

Sixteen volumes of this dictionary have been published (2008) spanning the Greek language from the letter *A* to the letter *Π* (entry word: *πνεύμονας*: = 'lung'). Prof. Kriaras (with his colleagues) has already collected the complete source entries of his work (letters *A* to *Ω*) which is under publishing process. The dictionary is considered to be one of the most important, renowned and complete lexicons of Greek language regarding the aforementioned period, as well as the total Byzantine period (Note 1). We shall refer also relative well-known works as:

(1) *Diccionario Griego-Español* edited by F.R. Adrados, (till 6th century A.D.) (Adrados & Somolinos, 1971-) [Madrid]

(2) *Lexikon zur byzantinischen Gräzität* by E. Trapp, (Trapp, 2001-) [Austrian Academy of Science]

(3) *Thesaurus Graecae Linguae (TLG),* (Estienne, 1831^1, 2002^2) [Paris]

(4) *Suidae Lexicon,* edited by A.Adler (Suidae, 1928-1938) [Leipzig]

(5) *A Patristic Greek Lexicon,* edited by G.W.H. Lampe*,* (Lampe, 1961-1968) [Oxford]

(6) *Greek Lexicon of the Roman and Byzantine Periods*, (Sophocles, 1887) [Cambridge, Massachusetts]

etc.

The basic sources of the Kriaras work are mainly texts of literature and history, while he also used as secondary sources theological and liturgical texts, poetry etc.

1.2 The corpus

We should initially state that we have to confine our processing to the already published volumes. Thus the resulting corpus is defined accordingly and it incorporates the entries of the first 16 volumes. We have developed a digital corpus, in matrix form, consisting of the above mentioned entries (lexemes) accompanied by their basic explanations of meaning. The characteristic problem regarding the whole work, especially in its electronic form, is the choice of the appropriate accentuation system. Although the original words were written using of course the Greek polytonic system, Prof.Kriaras declined to not use this system after the 5th volume (in 1977). The entries were digitized using in each case the accentuation system used by the original printed work. Finally we should also state that the lexicon is structured using *wordforms* and not *lemmata* (Jourafski & Martin, 2000).

1.3 The processing

The e-corpus resulting from the Kriaras lexicon is the hugest e-collection of the (historical) Greek language, if we should except the collection TLG (*Thesaurus Lingua Graecae*) (Note 2). Indeed, the dictionary refers to an extremely long period, while Kriaras attained to carry out an exhaustive work. We should state that the kind of processing performed using this corpus is the first endeavour ever done for the Greek language to such an extent.

We apply three fundamental types of language processing (Jourafski & Martin, 2000) regarding this corpus: (a) *metric measurements*, (b) *distributions* and (c) *spectra*. These kinds of transaction provide a metric-graphical analysis and imaging of the Greek language. For simplicity purposes we refer to case studies of sub-corpora in the lexicon of words beginning with the same letter.

(a) In the first case [*metric measurements*] we compute the mean word length of each letter (i.e. the mean length for all the words beginning with a specific letter) in the Greek alphabet.

(b) In the second case [*distributions*] the histogram of sub-corpora of individual initial letters, regarding the length of each entry (lexeme), is constructed. This feature is equal to the mapping of lexicon entries according to their *grammical length* (Note 3). Finally, we cite the normalized histogram for individual initial letters in the lexicon, where the lexemes are reordered (permuted) with the criterion of ascending order.

(c) By the term '*spectrum*' we incorporate graphs and data corresponding to rates and frequencies which determine the rhythms of (lexical) change inside the corpus. A characteristic and crucial spectrum of this category is the one referring to the lexical-length distribution. We name this field as the *lexical-length (l-l) spectrum or distribution,* or even, as the *IntraLexeme Frequency spectrum.* We cite for instance the case of the *l-l spectrum* for the letter *'I'* in Fig.3 and we'll discuss it in next paragraph.

2. Results

The outcomes of our analysis, using the e-corpus of the $\overline{m}GrL$, are given by samples of the overall project (Myridis, 2006). The reasoning is based on the following three terms: (a) the complete work has not yet been published (i.e. the rest 4 volumes) (b) in order to avoid burden in the presentation and (c) in order not to exceed space limitation of presentation. In consequence with these terms we decided to demonstrate the results regarding (a) a rare letter (i.e. the letter *I* (203 entries)) and (b) a frequent letter (the letter *N* (599 entries)). A frame of the overall dataset for the letter *I* is given in Table 1, while in Fig.1(a) the corresponding histogram is illustrated. In Fig.1(b) the normalized histogram is given for the '*I*' case. We also display the regular and the normalized histogram for the letter *N* in Figs.2(a) & 2(b). Finally, we have computed the mean word length for words beginning with *I* as

$$l_I=8,364532$$

while the mean word length for '*N*' is

$$l_N=8,21644$$

We may observe that the 'rare' letter *I* forms longer words than the 'frequent' letter *N,* although the number of I-lexemes is equal to 1/3 of N-lexemes. And this may constitute a reason of frequency scaling between these letters throughout the Greek language. We also notice that the mean word length converges for the chosen letters. Relative outcomes as well as additional processing could result for the rest of the letters.

2.1 Spectra

The *lexical-length (l-l) spectrum* turns out to be a useful tool for the investigation of the Medieval Greek lexicon and language. Indeed it reveals many important aspects among which:

1. *the distribution of the lexemes according to their length.* We may observe (Fig.3) that for the *'I'* letter, for instance, this distribution follows a Gaussian norm.
2. *the standard deviation (variance) of this distribution*, regarding normalized (inherent) frequencies (or *probabilities*) of wordforms' length for each initial letter. [When we use the absolute rate of lexeme's length then the most appropriate term is 'frequency', while the term 'probability' may be used in the case of normalized-to-the-unit rate (as in Fig.3(a)).]
3. *the construction of new, intrinsic tools of intra-language analysis*, although this analysis refers to lexicographical data.

etc.

We expand now the aforementioned aspects (case study: lexemes beginning with the letter *'I'*). Initially, we shall observe and analyze the form of the *l-l spectrum* (Fig.3(a)). In this figure, the normalized frequencies per lexeme's length is depicted, i.e. the probability of occurrence of a specific lexeme's length inside the specific letter sub-corpus. We can easily observe that the *l-l spectrum* in the 'I' case follows a Gaussian-like distribution (Papoulis, 2002). However we attempt to fit the data with even better accuracy. Models following Rayleigh distribution (Fernández, 2000) or Weibull functions (Weibull, 1951) do not fit the data properly. Thus we propose the model of the following type

$$\widetilde{p} = \frac{(\mathrm{n}+1)!}{(\mathrm{n}-\mathrm{k})!\mathrm{k}!} \tag{1}$$

which is also used to add more parameters to experimental probability distributions (Papoulis, 2002). In our model *n* is chosen to be (as of course is the case) equal to 20, i.e. the maximum length anticipated for '*I'*. Generally speaking, *n* is the maximum expected word length for lexemes. *k* is the length of a lexeme. A graphical representation of this model application is given in Fig.3(b). The physical interpretation of the model ruled by eq.(1) is the following. We assume the maximum (expected and practical) length, *n*, for an arbitrary entry regarding the specific initial letter.

The probability introduced by eq.(1) may be interpreted as: '$\widetilde{p}$ is the probability of occurring a lexeme of length *k* whilst the maximum possible length is *n.*' The comparison of Figs. 3(a) & (b) reveals the close relation between the measured data and the herein proposed model. The vital gain of this process is the generation of an analytical form in order to describe the mechanism of lexemes arrangement (*taxonomy*) in the $\overline{m}GrL$, with an inherent physical interpretation. This fact may pilot to the computational manipulation of lexemes' generation in arbitrary (computational) lexica.

We should also finally notice that the function used (eq.(1)) constitutes an ideal and intelligent catalyst for the transition of the probability density function (p.d.f) from a strict (inflexible) uniform law (regarding lexeme's length) to a smoother one (that prescribed by eq.(1)). Thus, the aforementioned model defines the a posteriori probability of occurring a lexeme of length *k* given that the maximum anticipated length can be *n*. We present this a posteriori probability density (a-pos p.d.f) in what immediately follows.

We finally refer to the usefulness of the proposed model to the lexical spectrum with respect to the entry length, and additional information resulting from the whole corpus of language. In this case we correlate information drawn both from lexicon source as well as from language corpora. In order to achieve this goal we consider the case of the a posteriori probability density function (p.d.f.) of the initial uniform distribution of entries' length (regarding independent collections of same initial letter entries) updated by the given condition of the knowledge of the previous state, i.e. regarding the entries' length of preceding lexemes (words). We can observe three kinds of this type of a posteriori p.d.f (Note 4). It can be proved (see also (Papoulis, 2002)) that all these p.d.f.'s may be represented by functions of the form (Fig.5)

$$f(p)= \tilde{p}\, p^k q^{n-k} = \frac{(n+1)!}{(n-k)!k!} p^k q^{n-k} \tag{2a}$$

which is approximately a beta function ($\beta(n,k)$) (or alternatively a binomial distribution)

$$\beta(n,k) \tag{2b}$$

In the constructed model (eq.(2a)) we have assumed uniform distribution for p and q.

(a) *the joint sequential probability of lexemes' length* (i.e. the estimation of the probability of ($k+1$)-length lexemes when the probability of k-length lexemes is known). In this case, p in eq.(2a) stands for the probability of occurring an entry with k symbols (e.g. as in Fig.4), while q is the probability of the rest entries' lengths (for the same initial letter in both cases).

(b) *the lexicon-based sequential linguistic probability of words' length* (for sequential words beginning with the same letter), i.e. the probability of a ($k+1$)-length word occurrence in speech (language), with the assumption that the probability of a k-length word occurrence (beginning with the same letter) just before the ($k+1$)-length word, is equal to the normalized probability of lexeme's length inside the lexicon's sub-corpus of entries beginning with the same initial letter. p and q are defined as in the previous case (a), however they refer to the probability of word occurrence in the (spoken or written) language. The assumption that the probability of a k-length word occurrence may be substituted, for the sake of application, by the probability of a k-length wordform occurrence in the sub-corpus of an individual initial letter in the lexicon, maybe is the only adequate and feasible approximation which may take place in practice, when the vast and untraceable field of real language must be considered.

(c) Then N_{AB}-*1 intra-letter* distribution diagrams can be constructed in a similar way to those diagrams resulting from the lexicon-based probability in (b). (N_{AB} is the number of letters in the alphabet (equals to 24 for the $\overline{m}GrL$)). The only difference is that in each diagram we assume that the preceding k-length word begins with some of the rest N_{AB}-*1* letters. In this case, the comments related to the previous type of a posteriori p.d.f. also hold. However p refers to the probability of a k-length word beginning with a specific letter (e.g. *'I'*), while q stands for the probability of every other occurrence of length other than k beginning with a different letter than the predefined (i.e. other than *'I'* in this example; e.g. *'N'*). Thus, in the case of *I-N intra-letter* probability distribution (or p.d.f.), p should take values as in Fig.3(a), while q refers to the probabilities of Fig.4.

We should ad hoc notice the correlation of lexicography to the language processing, through the use of the (a posteriori) lexemes' p.d.f.s and the introduction of the intermediate functions (beta, binomial etc.) which are used in order to approximate or fit real data (that of measurements and analysis of lexemes).

We should finally observe the surprising norms of spectra regarding the probabilities of occurrence of specific lexemes' lengths, for the sample letters *I* (Fig.3(a)) and *N* (Fig.4): the distribution for both the cases follows a Gaussian type function. That is: in a non-computational constructed language (such as the $\overline{m}GrL$), where no mechanistic rules and laws were predefined, the inherent structure of individual lexical sub-corpora reveals perfect symmetry and well-shaped, analytic, descriptive and cybernetic norms. This fact concludes to the result that it is a language of inherent wisdom.

3. Analytical points.

It is indeed useful to quote a couple of interesting remarks resulting from the lexical processing of the $\overline{m}GrL$.

1. The lexemes participate equivalently into the corpus, what is evident for every dictionary. However, they do not convey equal information. The measure, of course, for the information conveyed by each word is given by

$$p(\lambda)\log_2 p(\lambda) \tag{3}$$

The *entropy of information* (Cover & Thomas, 1991) of the random variable *Λ* is

$$H(\Lambda) = -\sum_{\lambda} p(\lambda)\log_2 p(\lambda) \tag{4}$$

where we have denoted an arbitrary lexeme by λ. *p(λ)* is the probability of appearance of this lexeme and *Λ* is the random variable referring to the lexemes of the lexicon. Obviously, the information conveyed by each word depends on the assumed corpus.

2. The length of each word influences, to some extent, the possibility of appearance of this word in speech, texts etc.

3. Alternative laws for measuring the information conveyed by each word in lexicons should be invented, in order the information quantity of dictionary lexemes to be self-defined without need for reference to outer data corpora.

4. We are also comparatively interested in relative frequencies of letter occurrences regarding English language (for instance letter *I=7%* (one of the five most frequent letters in English) and *N=6,7%*) (Barnett, 2009)

However the previous frequencies do not correspond to the frequencies of words beginning with these letters (*I and N).* In this case, for the Greek Medieval Language we could estimate a *relative lexical frequency (r.l.f.)* based on autonomous measurements inside the Kriaras dictionary.

5. *Intra-letter sub-corpora* using specific criteria, as for instance the specific meaning of lexemes (Table 2). In Table 1 we cite a sub-corpus belonging to the letter I, which consists of 15 entries. The mean word length of this sub-corpus is

$$l_{s,I}=8,901$$

which is obviously greater than the mean lexeme's length for the letter *I.* We may conclude that the words of this sub-corpus drastically increase the mean word length for the specific letter. Further linguistic outcomes could perfectly result from subsequent and more sophisticated processing. We should also state that the specific sub-corpus corresponds to 25% of the overall letter *I* dataset. Consequently, this is a master sub-corpus for the considered letter. We may characterize this corpus as a *quadrant sub-corpus* for letter *I.*

6. A typical hierarchical order is evident in lexica, as well as in each language, according to its alphabet (i.e. the letter *I* precedes letter *N*). There is not however such a hierarchy in spoken or written corpora (words beginning with *N* are not preceded hierarchically, according to their appearance, by words beginning with *I*).

Insert Table 1 and Table 2 Here

4. Conclusion

We focused in this paper on the e-corpus of the Medieval Greek Language based on the renowned lexicon of Kriaras. The collection is one of the most complete and huge corpora drawn from the period under examination. The existence of e-corpora of this kind enables the deep and useful analysis of a language and of its structure. We evaluate such an analysis using the Kriaras e-corpus and we present it in this paper. The results of the analysis are useful and exclusively indicative for the Medieval Greek Language. Finally, this kind of lexical and language processing regarding the Greek language is initially inaugurated by this work, with a plethora of fruitful results presented herein or anticipated in the future. The analysis and methodology presented by this paper could be perfectly applied in any other language/lexicon.

References

Adrados F.R. and Somolinos J.R. (eds.). (1971-). *Diccionario Griego-Español.* Madrid.

Andriotis N.P. (1995). *History of Greek Language*. Thessaloniki: Institute of Modern Greek Studies. (in Greek)

Barnett St. (2009). *Quantum Information*. N.York: Oxford University Press.

Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory.* N.York: Wiley.

Estienne H. (1831^1, 2002^2). *Thesaurus Graecae Linguae (TLG).* Paris: C.B.Hase - G.R.L. de Sinner - Th. Fix - A.F. Didot - L.G. Dindorf.

Evans H. (edt.) (2004). *Byzantium: Faith and Power (1261-1557).* N.York: The Metropolitan Museum of Art.

Fernández A.J. (2000). Bayesian inference from type II doubly censored Rayleigh data. *Statistics & Probability Letters*, vol. 48(5), p.393-399.

Hartmann R.R.K & James G. (1998). *Dictionary of Lexicography.* London: Routledge.

Jurafsky D. and Martin J.H. (2000). *Speech and Language Processing.* N.Jersey: Prentice Hall.

Kriaras E. (1969-). *Dictionary of Medieval Greek Language (1100-1669)*. Thessaloniki: Greek Language Center, (in Greek).

Lampe G.W.H. (1961-1968). *A Patristic Greek Lexicon*. Oxford.

Myridis N.E. (2006). *The Information of* Observation. Thessaloniki: Kyriakidis Pbls., p.283.

Papoulis A. & Pillai S.U. (2002). *Probability, Random Variables and Stochastic Processes* (4th eds.). N. York: McGraw Hill.

Sophocles [E. Apostolides]. (1887^1, 1992^2). *Greek Lexicon of the Roman and Byzantine Periods.* Cambridge, Massachusetts.

Suidae Lexicon, edited by A.Adler (1928-1938). vol. I-V, Leipzig.

Trapp E.(et al.). (2001-). *Lexikon zur byzantinischen Gräzität besonders des 9.-12. Jahrhunderts, Wien.*

Weibu*ll, W. (1951). A statistical distribution function of wide applicability.* J. Appl. Mech.-Trans. ASME, 18 (3): 293–297.

Notes

Note 1. www.greek-language.gr [date accessed: 23.10.2010].

Note 2. Thesaurus Lingua Graecae (TLG): www.tlg.uci.edu [date accessed: 23.10.2010].

Note 3. By the term *grammical length* we denote the number of letters into a single word.

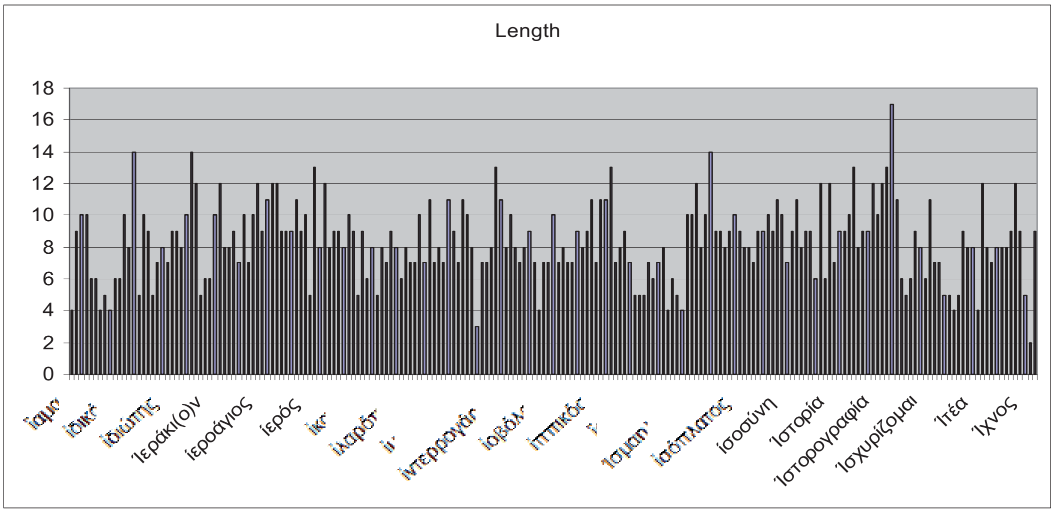
Note 4. The type of these probabilistic processing is different from *N-grams* (Jurafski & Martin, 2000).

Table 1. A frame of the sub-corpus of lexemes beginning with 'I'

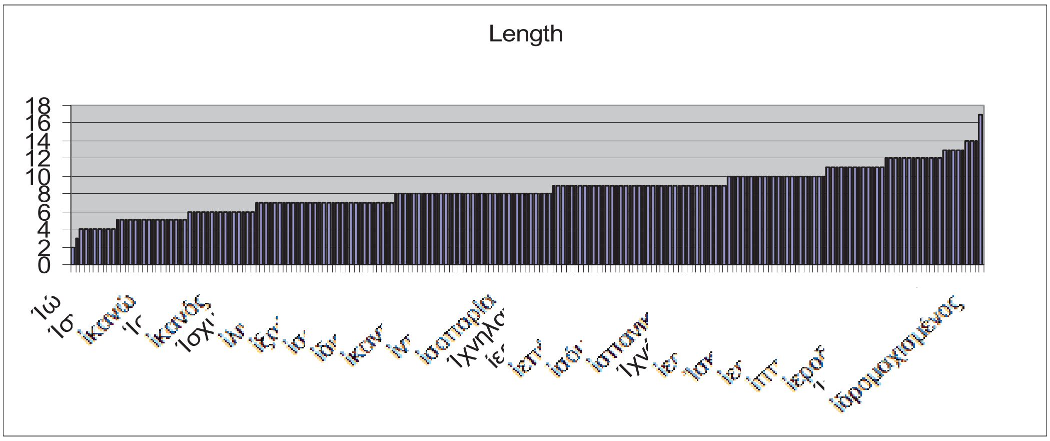
	Lexeme	**Explanation of meaning**	**Length**
1	ἲαμα	cure	*4*
2	ἰαματικός	who has the ability of curing	*9*
3	ἰαμβικάτος	something that is in iambic rhythm	*10*
4	ἰανουάριος	January	*10*
5	ἲασπος	jasper	*6*
6	ἰγδίον	mortar	*6*
7	ἰδέα	idea, shape	*4*
8	ἰδεῖν	seeing	*5*
9	ἲδια	similar	*4*
10	ἰδιάζω	to live alone	*6*
11	ἰδικός	faithful	*6*
12	ἰδιόγραφος	autograph	*10*
13	ἰδιοποιώ	arrogate	*8*
14	ἰδιοπροαιρέτως	voluntary	*14*
15	ἲδιος	same	*5*

Table 2. A frame of the sub-corpus of lexemes beginning with ‘I’: lexemes with philosophical meaning

	Lexeme	Explanation of meaning	Length
1	ἰδέα	idea, shape	*4*
2	ἰδεῖν	seeing	*5*
3	ἰδικός	faithful	*6*
4	ἰδιοπροαιρέτως	voluntary	*14*
5	ἱερός	sacred	*5*
6	ἱλαρότης	cheerfulness	*8*
	…….		

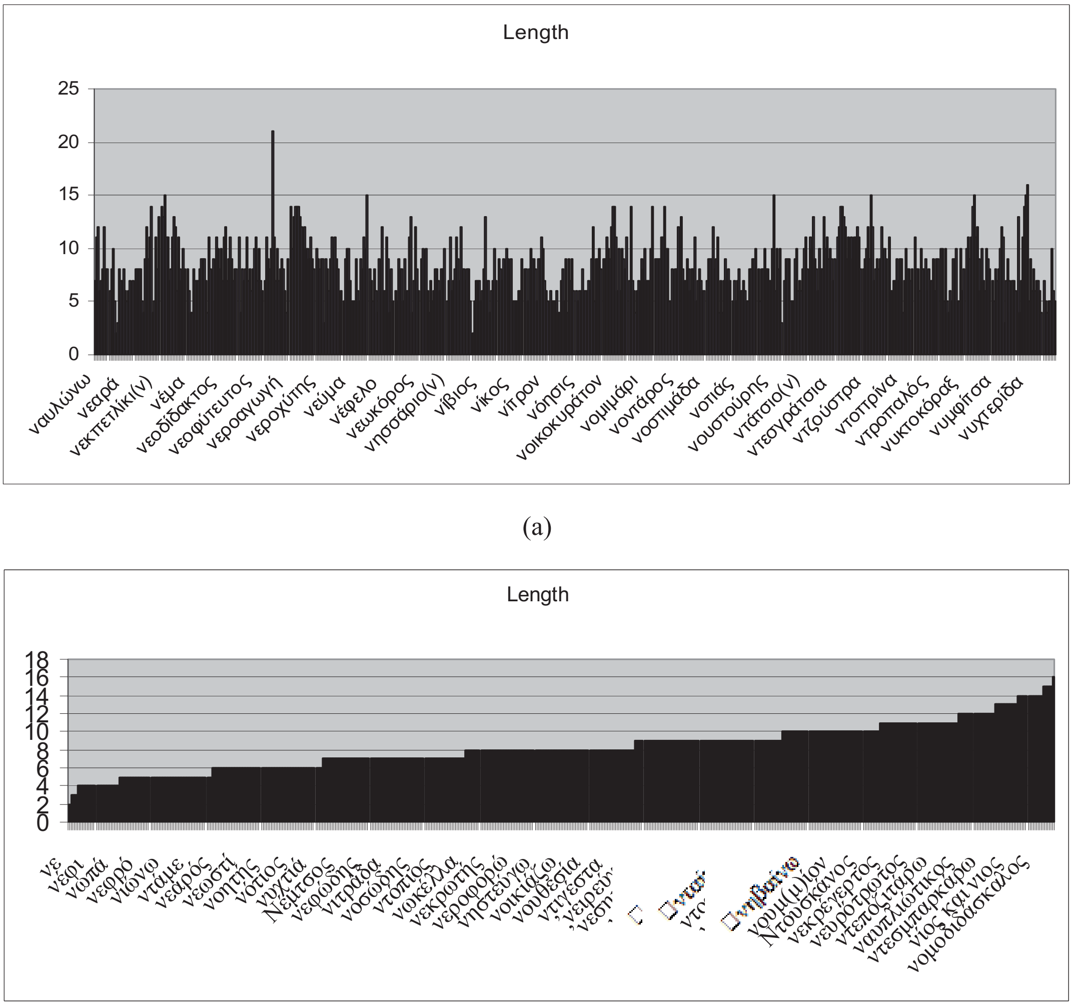


(a)



(b)

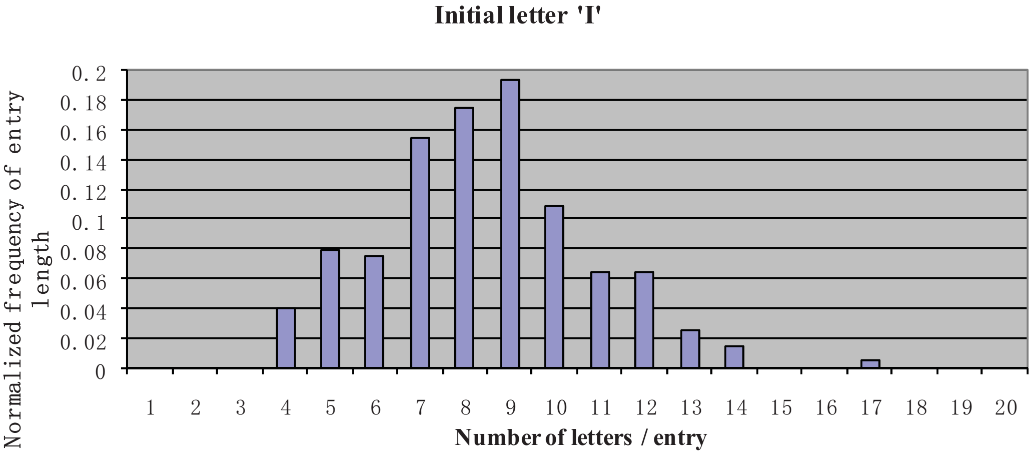
Figure 1. Words beginning with ‘I’ (a) histogram (b) normalized histogram



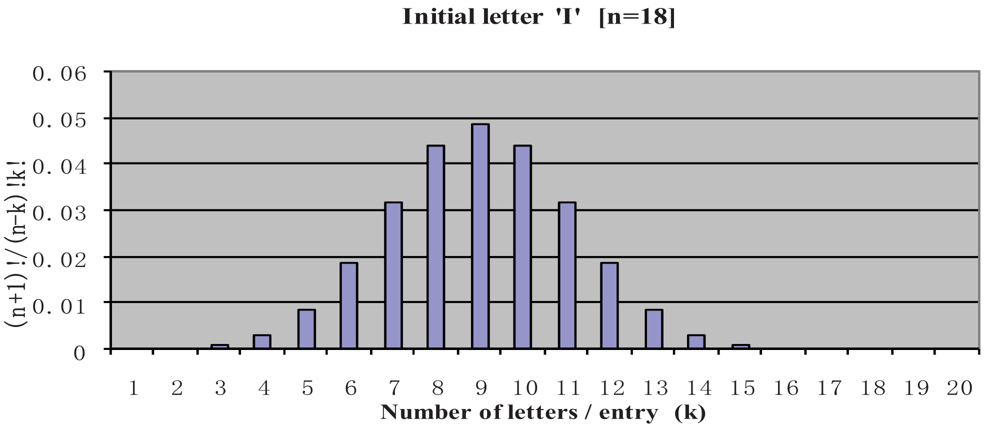
(a)

(b)

Figure 2. Words beginning with 'N' (a) histogram (b) normalized histogram



(a)



(b)

Figure 3. Sub-corpus of lexemes beginning with '*I*' (a) lexical-length (l-l) spectrum (b) the proposed model in eq.(1)

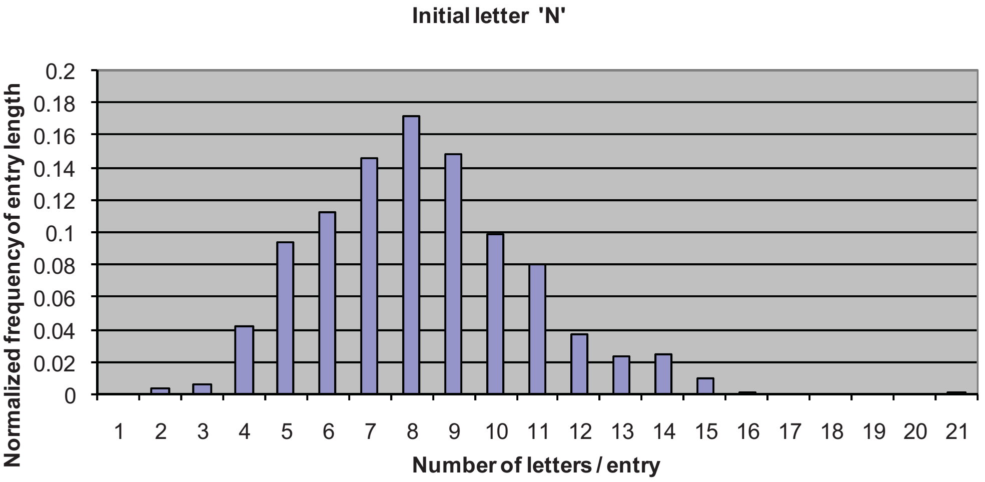


Figure 4. Sub-corpus of lexemes beginning with '*N*': the lexical-length (l-l) spectrum

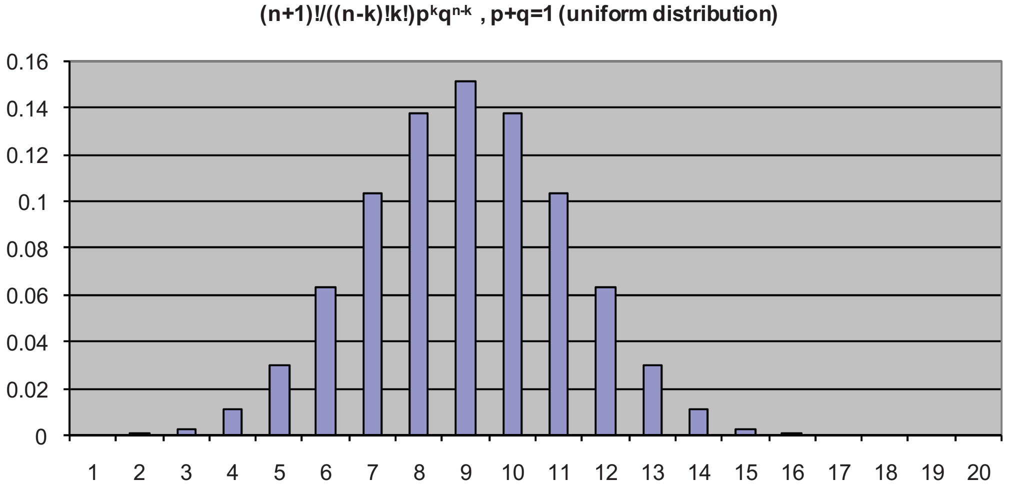


Figure 5. Lexical p.d.f. model based on eq.(2a)