Minimal Pairs and Functional Loads of Sound Contrasts Obtained from a List of Modern Greek Words

Constandinos Kalimeris, Stelios Bakamidis
Institute for Language and Speech Processing (ILSP), Epidavrou & Artemidos 6, Maroussi, 151 25, Athens, Greece
{c_kal, bakam}@ilsp.gr

Abstract
This paper reports on the initial results of our investigation into the distribution of speech sounds across the lexicon of Modern Greek (MG). The data we discuss ultimately derive from the list of orthographic word-types of a large general corpus of written MG. The orthographic word-types were automatically transcribed into their respective citation forms. Minimal pairs were automatically extracted from the resultant list of citation forms. The Functional Load (FL) of each sound opposition was computed as a function of (a) the length of citation forms, (b) the position of each sound contrast within citation forms and (c) the number of minimal pairs pertinent to each opposition question. The body of data yielded by this study will be used for further research in MG phonology as well as for the improvement of the performance of Automatic Speech Recognition applications.

Index Terms: minimal pair, functional load, contrastive distribution, phoneme inventory, Modern Greek

1. Introduction
The research work reported herein was originally undertaken in the hope of facilitating phonological investigations of Modern Greek (MG) [1]. Theoretical phonological analyses of MG abound but not uncommonly reflect views based on sparse data ([2] – [7]). To the best of our knowledge, a comprehensive body of data capable of supporting robust (i.e. data-driven) phonological analyses of MG has heretofore been non-existent. This paper reports on the building of such a body of data and attempts some tentative, yet data-driven, initial statements relevant to the phonology of MG.

The paper is organized as follows: Section 2 outlines a number of basic concepts. Section 3.1 discusses the study’s source (1st order) data. Section 3.2 discusses the study’s canonical allophonic (2nd order) data. Section 3.3 discusses and presents the combinatorial (3rd order) data obtained from the processing of the 2nd order data. Section 3.4 discusses and presents the Functional Loads (FLs) of pairs of sound segments in contrastive distribution across our wordlist and the FLs of candidate phonemes (4th order data). Section 4 lists and briefly discusses our initial phonological observations. Section 5 discusses our limitations and prospects for future research. Finally, Section 6 concludes the paper.

2. Basic concepts
Every written word can be said to have a typical phonetic realisation, a sort of mean which underlies all its instantiations in speech irrespective of inter- and intra-speaker variation within a given linguistic code. This representative phonetic mean may be called the word’s citation form or canonical form (cf. [8], [9]). It is commonly represented in writing as a string of phonetic symbols (representing discrete speech sounds, or sound segments, or phones) belonging to a special alphabet like the IPA [10]. The citation forms we use consist of the following set of 36 symbols: A = \{p, t, c, k, b, d, j, q, f, ð, x, v, ò, j, ŭ, s, z, ñ, m, n, ñ, r, l, ś, c, å, a, o, u, ĕ, i, á, ò, u\}. Set A is assumed to be sufficient for the representation of all citation forms relevant to the standard variety of MG.

The less predictable the occurrence of a symbol within the citation forms of a language, the more likely it is that the symbol is a phoneme of that language. The more predictable the occurrence of a sound symbol relevant to its neighbouring symbols is, the more likely it is that the symbol in question represents a conditioned phonetic variant (or allophone) of a ph
dene of that language. Ideally, phonemes are in contrastive distribution within canonical forms and allophones in complementary distribution (the terms are used in their linguistic sense [8]). It is obvious that the notions “phoneme” and “allophone” are not absolute but relevant to phone distribution within a particular body of data.

We take all symbols in set A (except the last 5) to be legitimate candidate phonemes of MG. The last 5 accented symbols are to be taken as shorthand notations. For example, [a] in a citation form denotes a phone with the segmental quality of an [a] which, additionally, happens to be the most predictable nuclear element of the syllable which bears the lexical stress of that citation form: [pata\textunderscore\textit{ta}]=\textit{[pa\textunderscore\textit{ta}]\ ('potato')}, [\textit{gaida}]=\textit{[gai\textunderscore\textit{da}]\ ('bagpipe')}, [\textit{triz\textunderscore\textit{dias}tato}]=\textit{[triz\textunderscore\omega\textit{ia\textunderscore\textit{sta}to]}\ ('3D')].

In this work, a pair of citation forms (i.e., a pair of strings of phonetic symbols, say \(w_1 = [a_1, a_2, a_3, \ldots, a_l]\) and \(w_2 = [b_1, b_2, b_3, \ldots, b_l]\)), is defined as a minimal pair iff:

\[
l_i = l_j = l
\]
where \(l_i\) and \(l_j\) are the respective lengths of \(w_i\) and \(w_j\), and

\[
a_i \neq b_i \quad \forall i \in [1, \ldots, k – 1] \cup [k + 1, \ldots, l],
a_i = b_i \quad i = k
\]
where \(a_i\) is the i-th character of string \(w_i\) and \(b_i\) is the i-th character of string \(w_j\). Note that, in this work, we consider all pairs of citation forms satisfying conditions (1) and (2) as legitimate minimal pairs, irrespective of the membership of each \(a_i\) and \(b_i\) in traditional linguistic categories such as “vowel” and “consonant”. This decision reflects our intention to produce FLs which will be as free from theoretical bias (i.e. as much data-driven) as possible. Hence, not only are \{\textit{arma}\}–\{\textit{alma}\\} (“tank”–“leap”), \{\textit{pınão}\}–\{\textit{ponao}\\} (“starve”–“hurt”) and \{\textit{stilos}\}–\{\textit{stilos}\\} (“pole”–“fleot”) considered legitimate minimal pairs, but also \{\textit{óasĩ}\}–\{\textit{ópsi}\\} (“oasis”–
The nature of our data (a wordlist, as opposed to a corpus) further forces us to adopt an even broader (linguistically-functionally speaking) view of the minimal pair: the members of a legitimate minimal pair cannot only correspond to different lexemes (as in the above examples), but also to different inflected forms of the same lexeme, or to forms of different lexemes which are closely related through morphological derivation. Hence, [trome]~[trote] (‘eatpres-1st-plural’ ~ ‘eatpres-2nd-plural’) and [enikos]~[enikus] (‘generally’~‘general_acc-plural’) are also considered legitimate minimal pairs.

3. Data

3.1. Orthographic (1st order) data

These derive from the list of word-types (unique orthographic words) of HNC [11]. Only strings consisting exclusively of Greek orthographic characters and bearing no more than one stress marks were retained. To achieve maximal source-data validity [1], the remaining orthographic forms were further filtered with the help of a morphological lexicon and a spelling-checker software [12], [13]. A stress diacritic was manually added to all non-clitic monosyllable forms corresponding to content words. The above procedures yielded a final list of about 210,000 orthographic word-types.

3.2. Allophonic (2nd order) data

The orthographic data were automatically transcribed into their respective canonical allophonic forms using an upgraded version of PHONEMIA [14]. The transcriber produced about 201,000 different canonical forms. The numerical difference across the 630 (theoretical) combinations of phones in contrastive distribution is a number of discrete objects (here, legitimate MG forms) which, according to (3), for $\mu = 7$ and $v = 2$, yields $\Sigma = 21$ minimal pairs:

$$\sum_{\nu}^\mu = \frac{\mu!}{v!(\mu-v)!}$$

where $\mu$ is a number of discrete objects (here, legitimate MG canonical forms) and $\Sigma$ is the number of possible non-ordered $v$-tuples (here, pairs) that these objects can form. [kobos], however, belongs to the 7-member paradigm [ko|br|ik|l|m|n|p] which, according to (3), for $\mu = 7$ and $v = 2$, yields $\Sigma = 21$ minimal pairs:

$$\sum_{\nu}^\mu = \frac{\mu!}{v!(\mu-v)!}$$

where $\mu$ is a number of discrete objects (here, legitimate MG canonical forms) and $\Sigma$ is the number of possible non-ordered $v$-tuples (here, pairs) that these objects can form. [kobos] is jointly responsible for 6 of the 21 (29%) relevant minimal pairs.

3.3. Minimal Pairs (3rd order data)

Around 149,500 minimal pairs were automatically extracted from the allophonic data. These are distributed unevenly across the 630 (theoretical) combinations of phones in contrastive distribution. The distribution of numbers of minimal pairs per sound contrast is given in Table 1a.

3.4. Functional Loads (4th order data)

A comprehensive review of the literature relevant to the quantification of the notion of Functional Load (FL) can be found in [15]. For our purposes, we adopt the following view of FL: the FL of a certain sound contrast operating within a lexicon must reflect the portion of that lexicon’s (quantifiable) ability to convey information through utilization of that contrast. Thus, the contrast [p~t] operating in the MG lexicon serves, among other things, to keep utterances like [pino] (‘drink’) and [tino] (‘tend’) apart. If MG stopped making that distinction, the words corresponding to such pairs of spoken signals would begin to sound the same, i.e., they would become homophones.

In this work, we define the functional load $FL[x~y]_{w,v}$ of two candidate phonemes $x$ and $y$ in contrastive distribution within two specific canonical forms $w_i = [a_1, a_2, a_3, ..., a_l]$ and $w_j = [b_1, b_2, b_3, ..., b_k]$ (i.e. within a specific minimal pair) as follows:

$$FL[x~y]_{w,v} = k / l$$

where $l$ is the length of either $w_i$ or $w_j$ and $k$ is the distance of the site of the sound contrast (measured in number of sound symbols) from the beginning of either $w_i$ or $w_j$. Hence, the $FL[x~y]_{w,v}$, with $x = [k]$ and $w_i = [\$a\$z\$m\$s\$]$ (‘silence’), and $y = [p]$ and $w_j = [\$p\$z\$m\$s\$]$ (‘spasm’), is $2 / 7 = 0.286$. By contrast, $FL[x~y]_{w,v}$, with $x = [k]$ and $w_i = [\$a\$n\$k\$s\$]$ (‘gathering’), and with $y = [p]$ and $w_j = [\$a\$n\$p\$s\$]$ (‘snakeps’), is $5 / 7 = 0.714$. The fact that the FL of the [k~p] contrast in the second pair is greater than the one in the first makes good sense: in the second case, the sound string [sina...], which needs to be retained in memory before it is disambiguated by the occurrence of either [k] or [p], is longer than the respective string [s...] in the first case. It is conceivable that an FL notion comparable to the one defined herein could be used to improve the performance of Automatic Speech Recognition (ASR) applications.

The functional load $FL[x~y]$ of a contrast between candidate phonemes $x$ and $y$ is defined as the sum of the FLs of every minimal pair pertinent to that contrast divided (i.e., normalised) by the sum of the FLs of all minimal pairs pertinent to all combinations of all candidate phonemes:

$$FL[x~y] = \frac{\sum_{r<s} FL[x~y]_{w,v}}{\sum_{r<s} FL[r~s]_{w,v}}$$

Finally, we may define the functional load $FL[x]$ of a candidate phoneme $x$ as the sum of all functional loads pertaining to each contrast of that phoneme and each of all the other members of A:

$$FL[x] = \sum_{r<s} FL[x~r] + FL[x~s]$$

A notion such as “the FL of a single phoneme” (as opposed to the FL of a contrast between two phonemes) is not self-contradictory, since phonemes are theoretical constructs which, by virtue of their very definition, are inherently contrastive entities [8], [9], [16].

The FLs of the various oppositions between candidate phonemes can be found in Table 1b. The FLs of the candidate phonemes of set A (symbols 1 – 31) are outlined in Figure 1.
4. Initial Observations and Discussion

Our 3rd and 4th order data lend themselves to sophisticated statistical analysis. This will be pursued in a future publication. For lack of space here, it will suffice to make a comparative study of the distribution of [s] within words in a number of languages may help to explain the long acknowledged problematic behaviour of [s] in the universal sonority hierarchy or sonority scale of sounds [9], [16]. The potential of [s] functioning (even if only marginally) as syllabic nucleus in the lexical and phrasal phonology of MG certainly warrants further exploration (c.f. [\'ko.\ddot{u}.ks] > [\'ko.\ddot{u}.kas] ‘code’, [\'pi.ks\'\ddot{u}.ak.s] ‘with punches and kicks’).

The 30 most frequent sound contrasts of the 446 attested in our data (Tables 1a and 1b) involve minimal pairs consisting either of two inflectional types of the same lexeme, or of types closely related through the process of morphological derivation, or of two type-variants related to MG diglossia (e.g. [\'x\ddot{a}.es] ~ [\'x\ddot{e}.es], ‘yesterday’) [5], [6].
This is reflected in the arrangement of Figure 1: the first 11 candidate phonemes are involved in such pairs. The distributions of minimal pair frequencies and of the FLs of candidate phonemes are likely to change dramatically if such pairs are filtered out. A first attempt towards this direction will be reported in a forthcoming paper.

Voiced plosives appear to be the least exploited sound class in MG, with the exception of [d], which participates in a large number of inflexional pairs. Their phonemic status is questionable and requires further investigation.

For [c] and [j], which are traditionally considered to be allophones of the phonemes /s/ and /l/ respectively, it holds that FL_{[c]} = FL_{[s]} and FL_{[j]} > FL_{[l]}. This is almost certainly due to the CV problem, treated in [4], [14] and, in greater detail, in [1]: instances of [c] and [j] in MG can also occur as allophones of the phoneme /l/.

5. Present limitations and future prospects

As alluded above, minimal pairs may not be of equal importance if viewed across the entire linguistic system of MG and not within the relatively narrow bounds of its lexicon or, indeed, of its phonological sub-component. Our future plans include the indexing and subsequent categorization of minimal pairs according to the lexicographical and morphosyntactic characteristics of their lexical members (e.g. part-of-speech, number, gender, case, person, tense, aspect, mood, etc.) and, following that, the re-computation of the FLs of the various sound contrasts and candidate phonemes.

Furthermore, there are plans for the exploitation of the data yielded by this study in the field of ASR applications.

6. Concluding Remarks

No final conclusions can be drawn on the basis of the data presented herein prior to their thorough analysis and subsequent re-categorisation: we hope that this study will only be the first in a series of future investigations in the distribution of speech sounds in MG. If anything, we believe that this study represents a promising first step towards the conduct of more robust (i.e. data-driven) phonological analyses of MG.

7. Acknowledgements

The first author would like to thank Mr John Higgins [17] for his encouragement back in summer 2002.

8. References


