A Rule-based Hyphenator for Modern Greek

Theodora I. Noussia*
Computer Technology Institute

The purpose of this paper is to formally examine hyphenation as it pertains to Modern Greek with the aim of achieving accurate and thorough machine hyphenation. Grammar rules are interpreted and formally expressed in terms of regular expressions of word substrings, and exact hyphenation rules are derived. Vowel splitting, which traditionally is indicated in terms of prohibitive rather than explicit grammar rules, is examined in detail. Many ambiguities caused by circular definitions of the prohibitive rule vowel sequences are detected, an overwhelming majority of which are resolved within the present framework.

1. Introduction

Hyphenator programs in modern typesetting systems are necessary to eliminate excess space between adjacent words in texts. Word hyphenation could be bypassed by stretching out this space, but this would effect the appearance of the document. A hyphenator program takes as input a word and returns the set of points within the word where hyphens are permissible. Word hyphenation depends strictly on the target natural language, and many of the problems encountered are language specific.

In general, machine hyphenation can be achieved either by consulting lists of hyphenated words or by developing pattern-based hyphenation programs (Liang 1983; Knuth 1986). The first approach ensures complete and correct hyphenation, but it has the disadvantage of being incapable of hyphenating words not on the list. In particular, for highly inflectional languages, such as Greek, these word lists would have to be extremely extensive in order to include all possible inflectional and derivational word forms. Even if such lists could be generated, it would be impossible to include words such as compounds, which can be readily created, or all proper names. In addition, the initial step toward the development of lists of hyphenated words is commonly rule-based hyphenation. On the other hand, although the second approach does not raise such problems, it has the disadvantage of being unable to guarantee complete and accurate hyphenation.

The aim of the present study has been to analytically examine Modern Greek hyphenation in order to develop a pattern-based hyphenator. The requirement specifications are defined as follows: (i) to strictly prohibit impermissible hyphen generation; (ii) to generate a hyphen list that is as complete as possible.

Existing hyphenator programs meet the first requirement either by decreasing the number of proposed hyphens or by establishing stop lists containing the appropriately hyphenated exceptional words. Commonly, fulfilling the second requirement depends on the development of extensive subword patterns associated with hyphenation rules, as in Liang 1983, for example. Establishing lists of exceptions has the same disadvantages as the approach to hyphenating through consulting lists of hyphenated words,

* Computer Technology Institute, 3, Kolokotroni str., 26 223 Patras, Greece. E-mail: noussia@cti.gr

and thus hyphenator dependencies on lists of exceptions must be restricted as much as possible.

Native Greek speakers are able to hyphenate most Greek words fully and unambiguously. In extreme cases, they will propose two hyphen sets for the same word, one being a proper subset of the other, but both being acceptable. However, complete automatic hyphenation is a rather complex task. Although consonant splitting is clearly determined by the grammar rules of Modern Greek and is thus easily expressed in terms of non-exceptional formal patterns associated with specific hyphenation rules, vowel splitting is not. The main problem of vowel splitting is that the grammar indicates the cases where splitting is not allowed, and the splitting of a large number of these cases is ambiguous. In addition, Greek vowels are sometimes accented, so ambiguity resolution concerns thousands of vowel sequences.

Existing hyphenator programs for Modern Greek are available as either commercial or research-based products and usually work on a minimal basis, i.e., finding only hyphenation points of consonant sequences. Some research-based versions, including one application of the TEX (Knuth 1986) hyphenator for Greek, achieve improved hyphenation but cover a minimal subset of the vowel sequences.

The present paper will encapsulate the standard grammar hyphenation rules and the general principles used in this study. The paper expresses these rules (which focus mainly on consonant sequences) formally and points out their limitations in terms of formal word expressions that can be completely and correctly hyphenated. The paper then turns to the problem of vowel splitting, and, by formally examining prohibitive grammar rules, deduces general hyphenation rules. It presents additional heuristic rules discovered during an exhaustive search of ambiguous vowel patterns, and demonstrates the degree of the resolved ambiguity in terms of the number of vowel sequences that have been disambiguated. Implementation issues are discussed, as well as the problem of words written in uppercase letters. Finally, the paper outlines the potential for generalization to other languages.

2. Hyphenation Rules

2.1 Consonant Splitting

According to KEME (1983)[1], the splitting of a Modern Greek word into syllables is governed by the following rules:

C1. A single consonant between two vowels is hyphenated with the succeeding vowel.

C2. A sequence of two consonants between two vowels is hyphenated with the succeeding vowel, if a Greek word exists that begins with such a consonant sequence. Otherwise the sequence is split into two syllables.

C3. A sequence of three or more consonants between two vowels is hyphenated with the succeeding vowel, if a Greek word exists that begins with the sequence of the first two consonants. Otherwise it splits; the first consonant being hyphenated with the preceding vowel.

The output of a hyphenator program is a set of permissible hyphen points within the input word. In order to specify this set, we shall proceed to a formal interpretation

1 This book is the official grammar book of Modern Greek edited by a group of experts and it is a revised edition of Triantafillidis (1941, reprint with corrections 1978).

of the grammar rules. As can easily be observed, the grammar rules are pattern based. Thus, the input word is divided into substrings, and the corresponding rules are applied to the substrings. Specifically, the goal is to identify the regular expressions of the patterns and the exact hyphen points for each formal pattern. Ambiguity issues caused by the interpretation of the grammar rules will be resolved. We will also prove that rules C1-C3 are not sufficient to provide complete hyphenation coverage of Greek words. This study will be based on rules C1-C3 and, in addition, the informal definition of a syllable as consisting of at least one or more vowels, or vowel(s) accompanied with one or more consonants (Triantafillidis 1978, 38), shall be adopted.

Let V be the set of vowel characters, C the set of consonant characters, $v \in V$, and $c \in C$. Specifically, $V = \{\alpha, \varepsilon, \eta, \iota, o, \upsilon, \omega, \acute{\alpha}, \acute{\varepsilon}, \acute{\eta}, \acute{\iota}, \acute{o}, \acute{\upsilon}, \acute{\omega}, \ddot{\iota}, \ddot{\upsilon}, \tilde{\iota}, \tilde{\upsilon}\}$, $C = \{\beta, \gamma, \delta, \zeta, \theta, \kappa, \lambda, \mu, \nu, \xi, \pi, \rho, \sigma, \tau, \varphi, \chi, \psi\}$. Subscripts, e.g., v_1, v_2, c_1, c_2 are used in order to distinguish more than one vowel or consonant of the same pattern. The beginning or end of the input word is indicated by the symbol "•". Optionality is indicated by placing characters inside square brackets. The operation obtaining one or more strings is denoted by the symbol "+". The operation obtaining zero or more strings is denoted by "*", or Kleene star (Lewis and Papadimitriou 1981).

We shall begin with the formal representation of the grammar rule subword patterns. The substrings of rules C1, C2, and C3 constitute one or more consonants between two vowels, or the strings of the expression $v_1c^+v_2$. Let c_1 be the first (obligatory) consonant in the consonant sequence of that expression. Let c_2 and c_3 be the second and the last (optional) consonants of the same sequence. Thus, the expression can be written as $v_1c_1[c_2c^*c_3]v_2$. Therefore:

Lemma 1
The substrings of grammar rules C1, C2 and C3 are contained in the set of the expression $v_1c_1[c_2c^*c_3]v_2$.

Grammar rules C1, C2, and C3 determine the hyphenation of word substrings comprising embedded consonants between vowels. They do not apply to substrings containing initial or final consonants. According to the informal definition of syllable given above, a syllable has at least one vowel and thus the consonant prefixes and suffixes of a word cannot constitute entire syllables. In other words, the maximal consonant prefix of a word is always hyphenated with the following vowel and the maximal consonant suffix of a word is always hyphenated with the preceding vowel. The permissible hyphen points of words are located between the syllables, thus:

Lemma 2
(a) The point following the maximal consonant prefix of a word and (b) the point preceding the maximal consonant suffix of a word do not constitute permissible hyphen points.

The substrings of lemma 2(a) comprise the set of all maximal prefix and consonant sequences of words. Formally, the set of expression $\bullet c_1[c_2c^*c_3]$ is the set of all maximal prefixes of consonants. Respectively, $c_1[c_2c^*c_3]\bullet$ is the expression for the set of substrings of lemma 2(b). Thus:

Lemma 3
The consonant substrings of Lemma 2(a) and 2(b) are contained in the sets of expressions $\bullet c_1[c_2c^*c_3]$ and $c_1[c_2c^*c_3]\bullet$, respectively.

Table 1
Consonant patterns and hyphenation rules. ($C = \{\beta, \gamma, \delta, \zeta, \theta, \kappa, \lambda, \mu, \nu, \xi, \pi, \rho, \sigma, \varsigma, \tau, \varphi, \chi, \psi\}$, $CC = \{\beta\gamma, \beta\delta, \beta\lambda, \beta\rho, \gamma\delta, \gamma\kappa, \gamma\lambda, \gamma\nu, \gamma\rho, \delta\rho, \theta\lambda, \theta\nu, \theta\rho, \kappa\lambda, \kappa\nu, \kappa\rho, \kappa\tau, \mu\nu, \mu\pi, \nu\tau, \pi\lambda, \pi\nu, \pi\tau, \sigma\beta, \sigma\gamma, \sigma\kappa, \sigma\mu, \sigma\pi, \sigma\tau, \sigma\varphi, \sigma\chi, \tau\zeta, \tau\mu, \tau\rho, \tau\sigma, \varphi\theta, \varphi\tau, \varphi\lambda, \varphi\rho, \chi\theta, \chi\tau, \chi\lambda, \chi\nu, \chi\rho\}$)

	Pattern	Condition	Hyphenation
c_1	$v_1c_1[c_2c^*c_3]v_2$	$c_1[c_2] \in CC \cup C$	v_1-$c_1[c_2c^*c_3]v_2$
c_2	$v_1c_1[c_2c^*c_3]v_2$	$c_1[c_2] \notin CC \cup C$	v_1c_1-$[c_2c^*c_3]v_2$

Now, let us formally specify the hyphen points as indicated by the grammar rules. According to lemmata 2 and 3:

Lemma 4
(a) For all words containing a substring $\bullet c_1[c_2c^*c_3]$, the point immediately following c_3, and (b) for all words containing a substring $c_1[c_2c^*c_3]\bullet$, the point immediately preceding c_1, are impermissible hyphen points.

In contrast, C1, C2, and C3 specify permissible hyphen points. However, two different interpretations can be given, namely that (i) only one hyphen point is specified by the rules, i.e., the point preceding or (exclusively) following the first embedded consonant c_1 or that (ii) two additional hyphen points are permissible: those preceding the first and following the second vowel. Both interpretations specify one *common* permissible hyphen point, which is, therefore, non-ambiguous. To define this point formally, let CC be the set of consonant sequences, two characters in length, that *can* begin a Greek word. (The exact definition[2] of set CC is given in Table 1. This set was extracted by the extensive listing of the initial word syllables presented in Setatos [1971]).

Theorem 1
The strings of the expression $v_1c_1[c_2c^*c_3]v_2$ are hyphenated as v_1 - $c_1[c_2c^*c_3]v_2$ if $c_1[c_2] \in CC \cup C$. Otherwise they are hyphenated as v_1c_1 - $[c_2c^*c_3]v_2$.

Proof
Rule C1 indicates that the strings of expression $v_1c_1v_2$ are always hyphenated as v_1-c_1v_2. These strings are a proper subset of $v_1c_1[c_2c^*c_3]v_2$ and they do not contain consonants $c_2c^*c_3$. Thus, $c_1[c_2]$ is degenerated to c_1, while $c_1 \in C$ by definition, and hence $v_1c_1[c_2c^*c_3]v_2$ are always hyphenated as v_1 - $c_1[c_2c^*c_3]v_2$.

The remaining strings are $v_1c_1c_2[c^*c_3]v_2$, and, as indicated by C2 and C3, their hyphen point is the point preceding c_1 if $c_1c_2 \in CC$ or the point between c_1 and c_2 otherwise. □

2 Some books state that three consonant sequences, namely $\mu\pi$, $\nu\tau$, and $\gamma\kappa$ (/b/, /d/, /g/) are excluded from set CC under specific contexts, see for example, paragraph 81, note 4 of Trantafillidis (1941) and paragraphs 140, and 141 of Tsopanakis (1994). The official grammar book (KEME 1983) however, does not treat these sequences as exceptional.

Theorem 2
The points immediately preceding v_1 and immediately following v_2 in the strings of expression $v_1c_1[c_2c^*c_3]v_2$ do not necessarily constitute permissible hyphen points.

Proof
Suppose that grammar rules indicate that the points immediately preceding v_1 and immediately following v_2 are also permissible hyphen points. By further taking into account Theorem 1, the assumption is that $v_1c_1[c_2c^*c_3]v_2$ is hyphenated either as - v_1 - $c_1[c_2c^*c_3]v_2$ -, or exclusively as - v_1c_1 - $[c_2c^*c_3]v_2$ -.

A hyphen is not permitted at the beginning or the end of the word, thus the possibility that the substring is located at the beginning or the end of the word is by definition excluded. Consequently, there is at least one character preceding v_1 and one following v_2. Consider the case where a consonant or consonant sequence precedes v_2. If it is at the beginning of the word, according to Lemma 4(a) the hyphen cannot be inserted after the consonant(s) and hence the assumption of a hyphen before v_1 is false. Respectively, for the case of a final consonant, or consonant sequence after v_2, according to Lemma 4(b) a hyphen following v_2 is not permitted. Now consider the case of a non-initial consonant or consonant sequence preceding v_1. In this case, Theorem 1 specifies one non-ambiguous hyphen point, which will not always be the point preceding v_1; hence the assumption is again, false. For the case of a nonfinal consonant sequence following v_2, the point implied by Theorem 1 may, in certain contexts, indicate the same point as the assumption, but in other contexts it may not. Nevertheless, in both cases the correct point will always be specified, thus the assumed rule does not need to be reapplied in order to indicate a potentially impermissible hyphen. Cases of a vowel preceding v_1 or following v_2 remain to be examined. In these cases, Theorem 1 does not define additional hyphen points, and Lemma 4 does not indicate impermissible hyphen points. The issue to examine is vowel splitting independent of consonants. As we shall see in the following section, vowel splitting is not always permissible. To present the proof in its entirety, it would be sufficient to give two contradictory examples, where the points preceding v_1 and following v_2 are not permissible hyphen points, e.g., *αυ-λή* [av-lí] 'courtyard', and *πα – λιός* [pa-liós] 'old'.
Therefore, the assumption does not always hold. □

A summarized formal definition of the hyphenation patterns and their associated rules as discussed above is presented in Table 1. Theorem 2 gives further support to the proposition that grammar rules are not capable of completely hyphenating all NL words.

Theorem 3
Rules presented in Table 1 are sufficient to completely hyphenate all words containing no consecutive vowels.

Proof
Every syllable has at least one vowel, thus a word cannot have syllables exceeding the number of its vowels, and it cannot have fewer syllables than the number of non-ending maximal consonant sequences. Let n be the number of vowels in a word not containing consecutive vowels. Then, if the word begins with a consonant or consonant sequence the number of non-ending maximal consonant sequences is n, or otherwise, $n-1$. Consequently, all such words have exactly n syllables. According to the definition of a hyphen, these words have exactly $n-1$ hyphen points.

According to Theorems 1 and 2, for each substring $v_1c_1[c_2c^*c_3]v_2$, precisely one hyphen point can be derived. All words containing n vowels, none of which are pairwise consecutive, have exactly $n - 1$ substrings of the expression $v_1c_1[c_2c^*c_3]v_2$, and according to Theorem 1, for each of these, one hyphen point can be derived. Therefore, for all words containing no consecutive vowels, precisely $n - 1$ hyphens are derived, and thus the rules of Table 1 are sufficient to completely hyphenate these words. □

2.2.1 Elimination of consonant sequences and loanword hyphenation. The examination of consonant splitting has not set any restrictions on the maximum length or even on the existence of certain consonant sequences. The phthong sequences that Modern Greek permits are, however, restricted by principles of grammar that are assumed to be universal. In the case of consonants, these principles state that the maximal consonant sequence is four characters long, and the maximal consonant prefix and suffix of Greek words are three characters and one character long, respectively (Setatos 1971). If these principles had been used in the examination of consonant splitting, the set of all subword patterns of Table 1 would have been restricted to the set of expression $v_1c_1[c_2c_3c_4]v_2$. Similarly, prefix and suffix consonant sequences would be restricted to $\bullet c_1[c_2c_3]$ and $c\bullet$, respectively. Furthermore, restrictions of specific sequences not possible in Greek words would further confine these patterns. Loanwords sometimes challenge these principles. Loanwords have been incorporated into Greek since ancient times and include words that cannot easily be recognized as borrowed, because of their adaptation into the above principles. Other loanwords, most frequently words that end in more than one consonant, e.g., φιλμ [film] 'film', have not completely adapted. A sequence of more than three consonants at the beginning of the word is also possible, as in the word Γκντανσκ [gdansk] 'Gdansk' (the city in Poland), although it is quite infrequent. Cases of more than four consonants may also exist or might appear in new loanwords or, most likely, occur in artificial words such as tongue twisters. Loanword hyphenation is governed by the same grammar rules as the rest of the language. Thus, in order to cover hyphenation of such loanwords, the patterns of Table 1 must not be eliminated. Apparently, this means that loanword hyphenation is independent of the rules governing hyphenation in the original language from which the word was borrowed. For example, although no hyphen point is derived by the Greek rules for the loanwords φιλμ [film] 'film', ταγκ [tank] 'tank', words having common derivatives of these, such as φιλμάκι [filmáki] 'small film' and τάνκερ [tánkɛr] 'tanker' are hyphenated as φιλ-μά-κι [fil-má-ki] and τάν-κερ [tán-kɛr].

2.2 Vowel Splitting

As already discussed, the rules presented in Table 1 cover hyphenation of word substrings containing at least one consonant; cases of vowel splitting are not covered. Vowel splitting is quite common in Modern Greek and is usually handled in grammar books with prohibitive rules. These are included within the context of the definition of various vowel combinations, but are rarely explicitly included within the set of standard hyphenation rules.[3]

Before proceeding to the presentation and analysis of these rules, some terms that will be used need to be defined. It should be noted that the terminology used refers to

3 Vowel sequences are sometimes explicitly mentioned in hyphenation rules, but usually only in the context of consonant sequences. For example, all references to vowels in rules about splitting of consonants may be augmented with "(or diphthongs)." This is not sufficient because vowel sequences that are not next to consonants may split, as in Παπα-ϊ-ω-άν-νου [papa-i-o-án-nou].

the orthographic representation of the various word substrings. Phonetic transcriptions are presented for the reader who is not familiar with Greek. Although phonetics is the ultimate basis for hyphenation, our approach is based on the available data, which is the orthographic representation of words, and not a transcription in a phonetic alphabet such as IPA.

As it has been previously defined, the term vowel refers to a single vowel or vowel character; V is the set of vowels. **Double-vowel blends** are phonetically equivalent to vowels, and their orthographic representation comprises two vowel characters. Let $2V$ be the set of double-vowel blends, $2V = \{\alpha\iota, \varepsilon\iota, o\iota, \upsilon\iota, o\upsilon, \acute{\alpha}\iota, \acute{\varepsilon}\iota, \acute{o}\iota, \acute{\upsilon}\iota, o\acute{\upsilon}\}$ ($\{[\varepsilon], [i], [i], [i], [u], [\acute{\varepsilon}], [\acute{\imath}], [\acute{\imath}], [\acute{u}]\}$). Some two-vowel orthographic combinations are phonetically equivalent to a vowel-consonant sound. Let VC be the set of such two-vowel combinations, $VC = \{\alpha\upsilon, \varepsilon\upsilon, \eta\upsilon, \acute{\alpha}\upsilon, \acute{\varepsilon}\upsilon, \acute{\eta}\upsilon\}$ ($\{[av], [ev], [iv], [\acute{a}v], [\acute{e}v], [\acute{\imath}v]\}$). Finally, **diphthongs** and **excessive diphthongs**[4] are vowel sequences consisting of two parts; each part can comprise either a vowel or a double-vowel blend. Precisely, the set of diphthongs and excessive diphthongs is a proper subset of $\{f_1f_2\colon f_1, f_2 \in V \cup 2V\}$.

The prohibitive hyphenation rules regarding vowel splitting are as follows:

V1. Double-vowel blends do not split.

V2. The combinations $\alpha\upsilon$, $\varepsilon\upsilon$, $\eta\upsilon$, $\acute{\alpha}\upsilon$, $\acute{\varepsilon}\upsilon$, and $\acute{\eta}\upsilon$ do not split.[5]

V3. Diphthongs do not split.

V4. Excessive diphthongs do not split.

All of the above rules are negative in that they indicate impermissible hyphen points within particular substrings of consecutive vowels. As the goal of the hyphenator is to identify the permissible hyphen points, we interpret V1, V2, V3, and V4 complementarily, i.e., in all other cases, splitting is allowed. It is important to note that the ultimate goal is to specify the permissible hyphen points in any vowel sequence, and not only in the particular substrings of sequences mentioned in V1, V2, V3, and V4. Formally, for every vowel sequence $v_0 \ldots v_{n-1}$ of n vowels, and its corresponding set of points $P_{v_0 \ldots v_{n-1}} = \{h_i\colon h_i \text{ is the point between } v_i \text{ and } v_{i+1}, 0 \leq i < n-1\}$, the issue is to identify set $IP_{v_0 \ldots v_{n-1}} \subseteq P_{v_0 \ldots v_{n-1}}$ of the impermissible hyphen points. Then the set $PP_{v_0 \ldots v_{n-1}}$ of the permissible points will be their set difference, or $PP_{v_0 \ldots v_{n-1}} = P_{v_0 \ldots v_{n-1}} - IP_{v_0 \ldots v_{n-1}}$.

Let us first formally specify the impermissible hyphen points in the particular sequences of V1–V4 rules. The combinations contained in V1, V2, V3, and V4 are distinguished in terms of their constituent elements. All combinations are made up of two parts; both parts of double-vowel blends and combinations $\alpha\upsilon$, $\varepsilon\upsilon$, etc. of rule V2 are vowels, while both parts of diphthongs and excessive diphthongs can be either vowels or double-vowel blends. Therefore, the impermissible hyphen point is located between the two parts in all combinations. For double-vowel blends and the elements of VC, which are digrams by definition, the impermissible hyphen point falls between its two constituent vowels, or

$$IP_{v_0v_1 \in 2V \cup VC} = \{h_0\} = P_{v_0v_1 \in 2V \cup VC} \text{ hence } PP_{v_0v_1 \in 2V \cup VC} = \varnothing. \tag{1}$$

Therefore, no additional hyphen point is derived for any word where each vowel sequence is of either the $2V$ or the VC type. Consequently, Theorem 3 is augmented

4 Diphthongs and excessive diphthongs will be defined operationally in the next pages.

5 The $\eta\upsilon$ combination is infrequently referred to in grammar books (KEME 1983), possibly because it appears in only a small number of words. However, this combination is also considered, because such words are regularly used e.g., $\varepsilon\varphi\eta\acute{\upsilon}\rho\alpha$ [efivra] 'I invented'.

to apply to words containing a maximum of two vowel substrings that are elements of $2V$ or VC.

Lemma 5
The rules presented in Table 1 are sufficient to completely hyphenate all words in which each vowel sequence is included in set $v_1[v_2]$, such that $v_1[v_2] \in V \cup 2V \cup VC$.

For words containing at least one n-gram vowel sequence, with $n > 2$, it is not apparent which vowel pairs, if any, will constitute a double-vowel blend or a VC so that the associated negative rules can be applied. Furthermore, diphthongs and excessive diphthongs comprised of either digrams consisting of two vowels, or trigrams consisting of a vowel and a double-vowel blend, or tetragrams consisting of two double-vowel blends, need to be precisely separated before the rules are applied. This procedure is called **tokenization** (see for example, Aho et al. [1986]). Tokenization in this case takes as input a vowel sequence and returns a sequential list of maximal non-overlapping tokens of the types $2V$, VC and V. Tokens *do not overlap* in that every vowel of the sequence is assigned to one and only one token. Tokenization might be ambiguous in that it might generate alternative token lists for specific vowel sequences. More precisely, alternative token lists can be generated for sequences where a vowel can be associated to its left or its right neighboring vowel in order to build up a $2V$ or VC token.[6] However, tokenization is achieved unambiguously because vowels are examined from left to right and a concrete token of the V type is extracted if it does not form a double-vowel blend or an element of the VC set with its subsequent vowel. Otherwise, a token of $2V$ or VC type is extracted. In conclusion, $2V$ and VC are disjoint, thus tokenization results in a unique list of tokens.

Let any vowel sequence $v_0 \dots v_{n-1}$ of n vowels, and its k-token sequence $f_0 \dots f_{k-1}$, $k \leq n$, $f_j \in V \cup 2V \cup VC$, $0 \leq j \leq k-1$, and let $P_{f_0 \dots f_{k-1}} = \{h_j \mid h_j$ is the point between f_j and f_{j+1}, $0 \leq j \leq k-1\}$. Let also $IP_{f_0 \dots f_{k-1}}$ and $PP_{f_0 \dots f_{k-1}}$ be the sets of impermissible and permissible hyphen points of the token sequence, respectively. Obviously, $PP_{f_0 \dots f_{k-1}} = PP_{v_0 \dots v_{n-1}}$ and $P_{v_0 \dots v_{n-1}} \supseteq P_{f_0 \dots f_{k-1}}$. According to (1), the elements of their set difference $P_{v_0 v_1 \dots v_{n-1}} - P_{f_0 f_1 \dots f_{k-1}}$ are all impermissible. Thus, the points that remain to be examined in regard to their hyphen permissibility are elements of set $P_{f_0 \dots f_{k-1}}$. This examination will be directed by V3 and V4 prohibitive rules. To conclude, formal definitions of diphthong and excessive diphthong sets would suffice. In this case, the specification of permissible hyphens would be based on whether each sequence of pairwise consecutive tokens is an element of one of these sets.

Identification of diphthongs and excessive diphthongs is a difficult task because of the ambiguity that arises when attempting to make specific designations. There are extreme cases where sequences exist whose assignment as diphthongs is context dependent. Some instances remain ambiguous even within precise contexts. Specifically, they may or may not be labeled as diphthongs, depending on the specific dialect or on the personal preference of the native speaker. To deal with this problem formally we shall determine weaker boundaries of diphthongs and excessive diphthongs. When considering hyphenation in regard to diphthongs, the problem is that diphthong definitions are circular, as in Triantafillidis (1978, 33), who states that "two vowels[7] that

6 As a matter of fact, the only sequences that might be problematic in tokenization are in the set of expression $\{\alpha \mid \varepsilon \mid \eta \mid o\}$ $\{\upsilon \mid \acute{\upsilon}\}$ $\{\iota \mid \acute{\iota}\}$. However, the algorithm ensures that the second vowel υ or $\acute{\upsilon}$ will be associated with the first. For example, the substring $o\upsilon\acute{\iota}$ in the word $\beta\varepsilon\delta o\upsilon\acute{\iota}\nu o\varsigma$ [veδuínos] 'Bedouin' is separated as $o\upsilon$ and $\acute{\iota}$ and not as o and $\upsilon\acute{\iota}$.

7 Double-vowel blends are included in this excerpt.

are pronounced together in one syllable constitute a diphthong." The circularity of this definition was observed by Petrounias (1984, 393), who proposes a less risky and more formal definition; namely, "a diphthong is made up of a semivowel and a vowel." Although the general issue of diphthongs is discussed at length, focusing on the development of ancient diphthongs and on phonetic issues, it is not discussed from a hyphenation perspective. A more clear approach is proposed by Tsopanakis (1994, 39, 84) who states that Modern Greek diphthongs are made up of a semivowel /i/ or /u/ and a vowel or double-vowel blend.[8] Although linguists do not always agree with the terminology, in order to avoid confusion, the simplified definition shall be used—namely that in diphthongs the semivowel is on the right and in excessive diphthongs, on the left.[9] Because semivowels are represented by vowel characters or double-vowel blends, the two categories overlap. In short, diphthong vowel sequences are suffixed with semivowels /i/ or /u/ and excessive diphthongs prefixed by an /i/ or /u/ semivowel.

The semivowel is not automatically identified, thus, practically, we have to deal with a large set of candidate diphthongs and candidate excessive diphthongs. Formally, these sets are defined as:

Candidate Diphthongs:	$CD = \{fx\colon f \in V \cup 2V,\ x \in I \cup U\}$
Candidate Excessive Diphthongs:	$CED = \{xf\colon f \in V \cup 2V,\ x \in I \cup U\}$

where I is the set of all vowels or double-vowel blends that are phonetically equivalent to /i/ and U is the set of all double-vowel blends that are phonetically equivalent to /u/.[10] The complete sets are as follows: $I = \{\eta, \iota, \upsilon, \acute{\eta}, \acute{\iota}, \acute{\upsilon}, \ddot{\iota}, \ddot{\upsilon}, \tilde{\iota}, \tilde{\upsilon}, \varepsilon\iota, o\iota, \upsilon\iota, \acute{\varepsilon\iota}, \acute{o\iota}, \acute{\upsilon\iota}\}$ and $U = \{ou, \acute{ou}\}$. Therefore, every sequence of two tokens f_jf_{j+1} in the k-token sequence $f_0 \ldots f_{k-1}$, $0 \leq j < k-1$, that is not a diphthong candidate or excessive diphthong candidate can, by definition, be split. Hence, formally the points between the tokens f_j and f_{j+1} of every sequence f_jf_{j+1} such that $f_jf_{j+1} \in (V \cup 2V \cup VC) \times (V \cup 2V \cup VC) - (V \cup 2V) \times (I \cup U) - (I \cup U) \times (V \cup 2V)$ are permissible. Equivalently, (i) the points adjacent to a VC token and any other token are permissible hyphen points and (ii) the points located between two consecutive tokens f_j, f_{j+1} such that f_j, $f_{j+1} \in (V \cup 2V) - (I \cup U)$ are also permissible points.

To conclude then, vowel splitting may be summarized by the following Lemma:

Lemma 6
For every vowel sequence f_1f_2, where f_1 and f_2 are maximal tokens, such that $f_1, f_2 \in V \cup 2V \cup VC$:

(a) If $f_1 \in VC$ or $f_2 \in VC$, the sequence is always hyphenated as f_1 - f_2.

(b) If $f_1 \in (V \cup 2V) - (I \cup U)$ and $f_1 \in (V \cup 2V) - (I \cup U)$, the sequence is always hyphenated as f_1 - f_2.

(c) Otherwise, the sequence is hyphenated as f_1 - f_2, if and only if, f_1f_2 is not a diphthong or an excessive diphthong.

Table 2 presents formal rules F2 and F3 for Lemma 6(a) and rule F1 for Lemma 6(b).

8 A character between slashes is an abbreviation for all different orthographical variations of the indicated phthong.
9 The use of this specific terminology does not affect the behavior of the hyphenator.
10 There are no vowel *characters* phonetically equivalent to the /u/ sound.

Table 2
Hyphenation rules of vowel patterns. ($V = \{\alpha, \varepsilon, \eta, \iota, o, \upsilon, \omega, \acute{\alpha}, \acute{\epsilon}, \acute{\eta}, \acute{\iota}, \acute{o}, \acute{\upsilon}, \acute{\omega}, \ddot{\iota}, \ddot{\upsilon}, \ddot{\iota}, \ddot{\upsilon}\}$, $I = \{\eta, \iota, \upsilon, \acute{\eta}, \acute{\iota}, \acute{\upsilon}, \ddot{\iota}, \ddot{\upsilon}, \ddot{\iota}, \ddot{\upsilon}, \varepsilon\iota, o\iota, \upsilon\iota, \acute{\varepsilon}\iota, \acute{o}\iota, \upsilon\acute{\iota}\}$, $U = \{o\upsilon, o\acute{\upsilon}\}$, $2V = \{\alpha\iota, \varepsilon\iota, o\iota, \upsilon\iota, o\upsilon, \alpha\acute{\iota}, \varepsilon\acute{\iota}, o\acute{\iota}, \upsilon\acute{\iota}, o\acute{\upsilon}\}$, $VC = \{\alpha\upsilon, \varepsilon\upsilon, \eta\upsilon, \alpha\acute{\upsilon}, \varepsilon\acute{\upsilon}, \eta\acute{\upsilon}\}$, $ID = \{\ddot{\iota}, \ddot{\iota}, \ddot{\upsilon}, \ddot{\upsilon}\}$, $IU_{\text{stressed}} = \{\acute{\eta}, \acute{\iota}, \acute{\upsilon}, \ddot{\iota}, \ddot{\upsilon}, \acute{\varepsilon}\acute{\iota}, o\acute{\iota}, \upsilon\acute{\iota}, o\acute{\upsilon}\}$, $ID1 = \{\ddot{\iota}, \ddot{\upsilon}, \ddot{\upsilon}\}$, $R = \{\rho\}$, $II = \{\iota\iota, \iota\acute{\iota}, \iota\eta, \iota\acute{\eta}, \upsilon\eta, \upsilon\acute{\eta}$, $\upsilon\varepsilon\acute{\iota}, o\iota\eta, o\iota\acute{\eta}, \iota\varepsilon\acute{\iota}\}$, $YI = \{\upsilon\iota\}$, $VY = \{\omega\acute{\upsilon}, \acute{\alpha}\upsilon, \iota\upsilon, \iota\acute{\upsilon}\}$)

	Pattern	Condition	Substitution
F_1	f_1f_2	$f_1, f_2 \in V \cup 2V - (I \cup U)$	f_1 - f_2
F_2	f_1f_2	$f_1 \in V \cup 2V \cup VC \wedge f_2 \in VC$	f_1 - f_2
F_3	f_1f_2	$f_1 \in VC \wedge f_2 \in V \cup 2V$	f_1 - f_2
F_4	f_1f_2	$f_1 \in IU_{\text{stressed}} \wedge f_2 \in V \cup 2V$	f_1 - f_2
F_5	f_1f_2	$f_2 \in V \cup 2V - (I \cup U) \wedge f_2 \in IU_{\text{stressed}}$	f_1 - f_2
F_6	f_1f_2	$f_1 \in V \cup 2V \wedge f_2 \in ID1$	f_1 - f_2
F_7	f_1f_2	$f_1 \in ID \wedge f_2 \in V \cup 2V$	f_1 - f_2
F_8	f_1f_2	$f_1f_2 \in VY$	f_1 - f_2
F_9	f_1f_2	$f_1 \in (V \cup 2V) - I \wedge f_2 \in (I \cup U) \cap 2V$	f_1 - f_2
F_{10}	f_1f_2	$f_1 \in U \wedge f_2 \in V \cup 2V$	f_1 - f_2
F_{11}	f_1f_2	$f_1f_2 \in II \vee (f_1 \in YI \wedge f_2 \in V \cup 2V)$	f_1 - f_2
F_{12}	$c_1c_2f_1f_2$	$f_1 \in I \cup U \wedge f_2 \in V \cup 2V \wedge c_1 \in C \wedge c_2 \in R$	$c_1c_2f_1$ - f_2

2.2.1 Diphthong Identification. In the previous section, vowel splitting was formally examined and concrete hyphenation rules were derived. However, as Lemma 6(c) explicitly acknowledges, hyphenation is restricted by diphthongs and excessive diphthongs. In this section, we shall proceed to an empirical examination of diphthongs and excessive diphthongs. Taking into account the initial specification that the hyphenator should never generate non-acceptable hyphens, and in order to pare down the enormous sets of candidate diphthongs and excessive diphthongs, we need to isolate the subset of sequences for which splitting is always permitted.

The approach followed was to first select all sequences of the above sets that were mentioned in various grammar books as examples of diphthongs and to assign them to the category of "neversplitting sequences." Then experimental matches were conducted through an electronic dictionary of Modern Greek that encodes 100,000 lemmata and all their inflectional and derivational forms (Vagelatos et al. 1995), and a 13 Mbyte corpus of newspaper articles. By definition, this process could not be automatic because hyphens were not included in the lexicon or the corpus, but there were far too many matches to be examined manually. Manual examination was restricted to those matches having limited frequency of occurrence. Nevertheless, during this process, a systematic method of identifying additional nonsplitting sequences was discovered based on a rule for stressing that states that a stress mark can only be applied to the ultimate, penultimate, or antepenultimate position of a word. Words in the lexicon were hyphenated based on the assumption that all remaining candidates *do* split. This hyphenation, however, resulted in certain words whose stress appeared on a syllable to the left of the antepenultimate position. Apparently then, incorrect hyphenation had been applied. All diphthong and excessive diphthong candidates included in these words were collected and designated nonsplitting sequences.

For the remaining candidates, identification of particular categories of substrings where a general exclusion rule may apply was attempted. Disparate and sometimes contradictory views given in various books (Setatos 1971; Triantafillidis 1978; Petrounias 1984; Mackridge 1987; Tsopanakis 1994) were collected. Their integrity was extensively examined through selection of matching words found in the corpus and in

the lexicon. This empirical process resulted in formally expressed rules independent of any exceptions. The sets of categories found are not necessarily disjoint, whereas all overlaps always lead to consistent hyphenation. All categories found are explained below, and representative hyphenated examples along with IPA transcriptions and translations are given. In order to avoid confusion, hyphenation is applied to those vowel sequences corresponding to the category currently being explained, and not to the entire word. Formal definitions of all categories are given in Table 2.

1. Examination of excessive diphthong candidates showed that 50% are immediately eliminated, i.e., always split. Specifically, rule F_4 states that candidate excessive diphthongs whose first part is stressed *do* always split, e.g., *παιδεί-α* [pεδí-a] 'education', *ιστορί-α* [istorí-a] 'history', *κύ-ησση* [kι-isi] 'pregnancy', *βίαιος* [ví-εos] 'violent', *λεί-ος* [lí-os] 'smooth', *Τροί-α* [trí-a] 'Troy'. On the other hand, not all diphthong candidates whose second part is stressed split, but the candidates in this set that are not simultaneously excessive *do* always split (rule F_5, Table 2).

2. Another category is associated with the existence of the diaeresis mark on a vowel of either a candidate diphthong or an excessive diphthong. All candidates whose second vowel has both a diaeresis mark and a stress mark *do* always split, e.g., *Μα-ΐου* [Ma-íu] 'May', *προ-ϋπαρξη* [pro-íparksi] 'preexistence', *εξα-ϋλωση* [εksa-ílosi] 'immateriality'. In addition, all candidates having as first or second token a *ϋ* always split, e.g., *ά-ϋλος* [á-ilos] 'immaterial', *προ-ϋπόθεση* [pro-ipóθεsi] 'prerequistic', *λαθροϋ-αλουργία* [laθroi-alurγía] 'glass smuggling'. As well, candidates that have as a first part only a nonstressed *ϊ* always split. Formally, this category is defined by rules F_6 and F_7 (Table 2). Diaeresis marks were used as a discriminating factor for additional candidates. The single-stress system imposed on Modern Greek in the last decade, states that "if the absence of the diaeresis mark does not generate ambiguity the mark should be eliminated" (Mackridge 1987, 93). Theoretically, this simplification could be applied to a variety of vowel sequences, but examination shows that acceptable words containing such sequences do not always exist, and not all sequences split. We focus on four that always split, namely: *ωύ - ζω-ύφιο* [zo-ífio] 'vermin', *άυ - ά-υλος* [á-ilos] 'incorporeal, immaterial', *ιυ - αρχι-υπηρέτης* [arxi-ipirétis] 'butler', *ιύ - περι-ύβρισση* [peri-ívrisi] 'insult' (rule F_8, Table 2).

3. Another observation is that all diphthong candidates having an *ου* or *ού* as a second part, and whose first part is not in set I always split. e.g., *κλαί-ουσα* [klé-usa] 'weeping willow', *ντα-ούλια* [da-úlia] 'drums', *μα-ούνα* [ma-úna] 'barge', *ωραί-ους* [oré-us] 'beautiful'. Furthermore, the category is expanded to include candidates whose second part is a double-vowel blend. At this point it should be stressed that there are specific examples where the candidate vowel sequence could linguistically be considered a diphthong based on pronunciation (Triantafillidis 1978, 19). However, during hyphenation they split de facto, e.g., *πά-ει* [pá-i] 'goes', *α-ειθαλής* [a-iθalís] 'evergreen' (rule F_9, Table 2).

4. Rule F_{10} is associated with those candidate excessive diphthongs that have an *ου* or *ού* as a first part. Note that the stressed /u/ has been

already included in F_4 because of the stress mark. Detailed examination of the candidates of this category led to the conclusion that the candidates always split during hyphenation. Although sometimes they are pronounced as diphthongs, they are split de facto, e.g., *Φεβρου-άριος* [fevru-ários] 'February', *βου-ητό* [vu-itó] 'clamor', *βου-ίζει* [vu-ízi] 'it clamors', *Βεδου-ίνος* [νεδu-ínos] 'Bedouin', *Ου-αλία* [u-alía] 'Wales', *ακου-ομετρία* [aku-ometría] 'acoustic metrics'.

5. An interesting subset of candidates concerns the intersection of candidate diphthong and excessive diphthong sets. This set is $(I \cup U) \times (I \cup U)$ and although it comprises a relatively great number of elements, most of these have low frequency of occurrence in linguistically acceptable words. It should be noted here that some parts of this set have already been covered by other rules. For the subset not covered, no general rule was formulated but particular instances that always split were identified. These instances are covered by rule F_{11}. We observed that some cases present ambiguity, while others always split e.g., *δι-ιστάμενος'* [δi-istámεnos] 'contrary', *δι-ίσταμαι* [δi-ístamε] 'I dissent', *δι-ηθημένος* [δi-iθiménos] 'filtered', *δι-ηπειρωτικός* [δi-ipirotikós] 'intercontinental', *δι-ήθηση* [δi-íθisi] 'filtering', *δι-ήγημα* [δi-íγima] 'short story', *μυ-ημένος* [mi-iménos] 'initiated', *μυ-ήσεις* [mi-ísis] 'initiate', *ποι-ητής* [pi-itís] 'poet', *ποι-ήσεις* [pi-ísis] 'you will do', *αυτοφυ-είς* [aftofi-ís] 'self-grown', *επιπλοποι-είς* [εpiplopi-ís] 'furniture-makers', *αλι-εία* [ali-ía] 'fishing', *υι-ός* [i-ós] 'son', *υι-οθεσία* [i-oθεsía] 'adoption', *άρπυι-α* [árpi-a] 'harpy'.

6. There is a different rule for determining the splitting of excessive diphthongs, referred to by both Triantafillidis (1978, 38) and Tsopanakis (1994, 108). It concerns the natural semantics of excessive diphthongs; the avoidance of hiatus in the spoken language. If the flow of speech is constrained by the existence of additional "difficult" or complex phthongs, the pronunciation of the excessive diphthong in one syllable becomes impossible. One such case is that of at least a double-consonant sequence, whose second consonant is *ρ* [r] followed by a candidate excessive diphthong. That diphthong is not excessive and should always be split (rule F_{12}, Table 2).

It should be noted that additional rules covering additional vowel sequences under specific contexts have been found and examined. For example, candidate diphthongs located between the members of compound words prefixed by a preposition do not split. The automatic identification of these instances would be based on a morphological analysis of words, a process beyond the scope of the present analysis.

2.2.2 Elimination of non-existent sequences. Having completed the vowel splitting study, the question of whether all sequences presented in the rules of Table 2 exist within *acceptable* Modern Greek words arises. Eliminations of consonant patterns exceeding a maximum length have already been discussed. Eliminations based on the existence of certain vowel sequences may be possible. However, ancient Greek words and borrowed foreign words that are frequently used in both written and spoken forms contain additional sequences and, as has already been mentioned, their hyphenation is governed by the same rules. Nevertheless, vowel sequences that contain consecutive stressed vowels or double-vowel blends, or consecutive vowels with diaeresis marks do not exist in any word—pure Greek or loan—and thus this can be used as a general

elimination principle. The patterns $f_1f_2, f_1, f_2 \in V \cup 2V \cup VC$ of Lemma 6 contain exactly 301 such sequences. From the remaining vowel sequences of Lemma 6, a few may be identified as non-existent. However, ad hoc compounds that can be readily created may contain even those sequences. Mackridge (1987) notes that, unlike with English, a person fluent in Greek has no difficulty in pronouncing an unknown word. This holds for all vowel sequences in Greek independently of whether they exist within acceptable words. It was thus decided to examine all theoretically possible cases and not to eliminate a priori any sequences.

2.3 Degree of Hyphenation Completeness

The rules in Tables 1 and 2 guarantee 100% correct hyphenation. The rules in Table 1 are capable of locating all permissible hyphenations of consonant sequences. In regard to vowel sequences, set $(V \cup 2V \cup VC)$ has 34 elements and according to Lemma 6 complete hyphenation of vowel sequences depends on $34^2 = 1{,}156$ vowel sequences. Grammar rules $V1$ and $V2$ explicitly define 16 of these, namely the elements of sets $2V$ and VC, while grammar books refer to 8 diphthongs that never split. Hence, only $16 + 8 = 24$ sequences were initially non-ambiguous, while $1{,}156 - 24 = 1{,}132$ were ambiguous. Rules F_1-F_{11} (Table 2) resolve the ambiguity of 1,015 different patterns. (Occurrences of overlapping patterns have been eliminated by analytically calculating the intersection of the sets of patterns for all pairs of rules F_1-F_{11}). In general, $1{,}156 - 24 - 1{,}015 = 117$ remain ambiguous. Thus, these rules are capable of completely hyphenating at least $(1{,}015 + 24/1{,}156)*100 = 89.9\%$ of the 1,156 sequences. (If non-existent patterns were eliminated, i.e., those consisting of either two consecutive stressed vowels or of two consecutive vowels with diaeresis marks, the degree of completeness of the hyphenator on a vowel pattern basis could be then computed as: $(1{,}029 - 301)/(1{,}156 - 301)*100 = 85.2\%$). Taking into account rule F12, which resolves ambiguity by proposing additional hyphen points under specific contexts, the degree of completeness increases. Furthermore, the ambiguity of additional sequences can be resolved without proposing additional hyphens, by using the rule stating that stress cannot be applied to a syllable beyond the antepenultimate position.

The degree of completeness calculated above does not represent completeness in terms of hyphenated words of real text corpora. The degree of complete hyphenated words of newspaper texts was manually calculated to be over 99%, as expected, because the frequency of occurrence of the remaining ambiguous vowel sequences in words of real texts is relatively low.

3. Implementation

In the previous sections, hyphenation issues were examined as they pertain to Modern Greek with the goal of achieving machine hyphenation that is both accurate and complete to the highest degree possible.

Existing hyphenators for Greek are commercial products and usually work on a minimal basis, i.e., finding the hyphen points of consonant sequences and, in limited cases, hyphens of vowel sequences. A research-based version of the Greek TEX typesetting system (Knuth 1986) provides improved hyphenation, but it only indicates splitting for 7.1% of the vowel sequences, which seem to have been selected rather intuitively. Furthermore, three of the sequences, as was observed, can generate impermissible hyphens.

The rules presented here have been used for the development of a hyphenator program included in the Microsoft Word for Windows 6.0 and 7.0 (Greek version) already on the market. The system has also been ported to different platforms including

Lotus AmiPro and a specialized typesetting system of a major Greek newspaper. The formal rules and the exact definitions of the sets of vowel and consonant sequences compiled in Tables 1 and 2 are sufficient to implement the hyphenator program. Patterns in Table 2 constitute maximal vowel tokens, which can be derived by a lexical analysis process, while patterns in Table 1 consist of single vowels and consonants.

The hyphenator program comprises two parts: the lexical analyzer and the actual hyphenator. The lexical analyzer reads the input characters and produces as output a sequence of maximal V, $2V$ and VC tokens, as well as tokens of the maximal consonant sequences of the word. For all tokens, the absolute starting position of the token in the input word is maintained, while the length of each token is implicitly defined by the token itself. All consonant tokens are also subdivided according to whether their two character prefix is contained in the CC set or not. Nontrivial consonant sequences are also designated by a flag indicating the occurrence of a ρ [r] suffix.

Vowel tokens are further classified according to the nearby resident vowel and consonant tokens. No additional classification of vowel tokens is needed in the following cases: (i) vowel tokens not in the $I \cup U$ set; (ii) vowel tokens that appear between any consonant sequences; (iii) stressed vowel tokens in the $I \cup U$ set that have as a left neighbor a consonant sequence with an /r/ suffix; (iv) vowel tokens that simultaneously have stress and diaeresis marks. The remaining vowel tokens are characterized explicitly as stressed I, nonstressed I, and U.

The actual hyphenation phase follows, where the hyphenator traverses the token sequence, identifies all ordered sequences of type (a) ⟨*vowel token*⟩ - ⟨*consonant token*⟩ - ⟨*vowel token*⟩, and (b) ⟨*vowel token*⟩ - ⟨*vowel token*⟩, and applies the corresponding hyphenation rules. The resulting hyphen points are given in terms of the absolute starting position in the word of the first or the second token of the sequence currently being examined.

3.1 Hyphenation of Words in Uppercase

There is no one-to-one correspondence between uppercase and lowercase letters. The main difference is that stress markings are not applied to words whose letters are all written in capitals while the diaeresis mark is maintained in capital letters.[11] Consequently, the transformation of any uppercase word to lowercase and back to uppercase again loses no information. The opposite transformation is not always without loss of information. To decrease the complexity of the hyphenator, we used only lowercase patterns. Thus, uppercase words are transformed to lowercase, hyphenated, and transformed back to uppercase forms.[12]

Hyphenation patterns of consonant sequences (Table 1) are unchanged because consonants do not take stress marks and, moreover, the vowels contained in these patterns are independent of stress. On the other hand, many of the patterns derived for the hyphenation of vowel sequences cannot be applied to capitalized words because the most important discriminating factor in diphthong identification is stress marking, and uppercase letters (Section 2.2.1) lack stress markings. This observation certainly implies the tendency for words in uppercase to have fewer hyphens than their lowercase equivalents. This inconsistency cannot be resolved without additional information about the position of the stress mark.

11 In words written with both capital and lowercase letters, an initial capital letter may have a stress mark.
12 The tranformation takes into account the existence of a final [s] in the uppercase word and tranforms it to the final ς instead of σ, according to a corresponding transformation rule.

4. Discussion

Overall, it was feasible to make an analytical examination of the hyphenating system mainly because most of the known hyphenation properties *were* expressed or *could be* expressed in terms of orthographic representation. In Greek, this representation contains much of the pronunciation information, which is the ultimate basis for hyphenation in every language. When analytical work reached the point where the available data could no longer provide the necessary pronunciation information, it was replaced by empirical work.

A similar process would be difficult to conceive in languages in which the orthography and pronunciation are significantly different. It should perhaps be stated that the system itself may not have the capacity to be generalized to other languages. It is interesting to note that rules governing the splitting of subword patterns exist in languages such as English, but their application is usually determined by orthographically inexplicit information, such as the existence of a long, short, or stressed vowel in some position of the pattern. Different types of properties typical of such languages as English and German are based on morphological considerations that were not an issue for our system. For example, in English "common roots" is an issue in hyphenation of compounds, whereas in Greek, it is not. Such properties are not likely to be similarly expressed in a pattern-based model. The process of developing a similarly performing hyphenator for such languages would be different. Identification of certain patterns would presumably be based on an empirical rather than an analytical process. Automatic extraction of common hyphenating properties from on-line hyphenated dictionaries is known (Liang 1983). The resulting patterns tend to be more detailed and extended. Lists of exceptions seem to be obligatory in such an approach because their lack would lead to the generation of impermissible hyphens.

5. Conclusions

Hyphenation issues pertaining to Modern Greek have been analyzed, and correct and thorough machine hyphenation has been achieved as a result of the present study. The explicit interpretation and formal expression of specific grammar rules led to a formal hyphenation model, and further provided a means of expressing the model's limitations. These limitations were in turn examined through an empirical process, which also resulted in formally expressed rules.

Acknowledgment

The author is grateful to M. Stamison-Atmatzidi for her long hours of proofreading, to the three CL reviewers for their valuable suggestions and comments, and to the Greek newspaper *To Vima* for the availability of the text corpus.

References

Aho, Alfred V., Ravi Sethi, and Jeffrey D. Ullman. 1986. *Compilers, Principles, Techniques, and Tools*. Addison-Wesley.

KEME. 1983. *Revision of Modern Greek Grammar of Manolis Triantafillidis* (in Greek). Didactic Books Publishing Organization.

Knuth, Donald. E. 1986. *The TEX Book*. Addison-Wesley.

Lewis, Harry and Christos Papadimitriou. 1981. *Elements of the Theory of Computation*. Prentice-Hall Software Series.

Liang, Frank M. 1983. *Word hy-phen-a-tion by computer*. Ph.D Thesis, Stanford University.

Mackridge, Peter. 1987. *The Modern Greek Language*. Oxford University Press.

Petrounias, Evangelos. 1984. *Modern Greek Grammar-Comparative Analysis. Volume A: General Linguistic Fundamentals, Phonetic, Introduction to Phonology, Part A: Theory* (in Greek). University Studio Press, Thessaloniki.

Setatos, Michalis. 1971. *Phonology of Modern Greek Koine* (in Greek). Papazisis Publishing, Athens.
Triantafillidis, Manolis. 1941. *Modern Greek Grammar (Dimotiki)* (in Greek). Reprint with corrections 1978. Institute of Modern Greek Studies, Thessaloniki.
Tsopanakis, Agapitos. 1994. *New Greek Grammar* (in Greek). Second Edition. Athens-Thessaloniki.
Vagelatos, Aristidis, Theodora Triantopoulou, Christos Tsalidis, and Dimitris Christodoulakis. 1995. Utilization of a Lexicon for Spelling Correction in Modern Greek, *10th Annual Symposium on Applied Computing—Special Track on Artificial Intelligence*, Nashville, TN, February.