Automatic Corpora-based Stemming in Greek

G. Tambouratzis and G. Carayannis
Institute for Language and Speech Processing, Greece

Abstract

In this paper, a system is presented that performs an automated morphological categorization of Greek words extracted from a corpus. This system processes morphologically the words via the repetitive application of a masking-and-matching technique. It is found that the introduction of *a priori* information regarding the grammar of the Greek language considerably improves the word segmentation accuracy. The system accuracy is evaluated by comparing the word segmentation with the entries of a morphological lexicon of the Greek language. The experimental results indicate that the output of the automated system is—for the majority of words—in agreement with the entries of the morphological lexicon. The proposed system is successfully applied to the generation of specialized morphological lexica on the basis of corpora consisting of term-intensive documents. Finally, possible extensions of the proposed system to other languages as well as to cover the derivation phenomenon for the Greek language are briefly reviewed.

1 Introduction

The Greek language is characterized by its highly inflectional nature. The continuous evolution of this language for over 3000 years means that although the origins of most lexical units (words) remain the same, their function in the inflectional system has changed. Thus, throughout the evolution of the Greek language it is the endings that have changed the most. For example, the endings in ancient Greek are considerably different from the corresponding endings in the modern Greek language, whereas the stems have a higher degree of similarity. Considerable differences also exist between the endings in formal Greek (Katharevousa)—used as the official language of the Greek state until 1979—and casual Greek (Dimotiki), although in the majority of cases the stems do not change. Hence, whenever a morphological lexicon needs to be created by hand, it is necessary to determine the time period of language evolution to which it refers. To encompass the language evolution through the ages, one would need to generate manually a large number of different morphological lexica, each covering a specific Language Evolution Sample (LES)

Correspondence:
G. Tambouratzis, Institute for Language and Speech Processing (ILSP), Artemidos & Epidavrou, Paradissos Amaroussiou, Athens 151 25, Greece.
E-mail:
giorg_t@ilsp.gr

corresponding to a particular time-period in the history of the Greek language and/or a geographical area. Consequently, significant benefits can be obtained if a method is devised for generating morphological information with the minimum possible human intervention. The research presented here is intended to provide tools useful for solving these issues and to allow the 'morphological processing' (and in particular the stemming) of a given set of words in an automated manner, with the minimum amount of both external guidance and post-processing.

The approach chosen needs to be fairly general to process texts from any LES with the minimum amount of modification. At the same time, it is desirable that it possesses the highest accuracy possible. Therefore, if some general characteristics of the LES are known (for example, some frequent endings or grammar rules), these may be provided to the system so as to optimize its performance. Evidently, there exists a trade-off between the requirement for the highest possible accuracy (which necessitates a large amount of *a priori* knowledge) and the requirement for the minimum external guidance (and thus the minimum amount of *a priori* knowledge). Different approaches (involving varying amounts of *a priori* knowledge) have been followed to investigate the system behaviour. To achieve that, the proposed Automated Morphological Processor (AMP) has a modular structure, allowing it to be tailored to a specific application environment by simply modifying the contents of some modules. The previous considerations are not of theoretical value only. For text retrieval and information extraction purposes when working with texts in the Greek language, it is important to be able to use two or three different stemming operators, to represent fully the inflectional evolution of the language through time.

2 AMP Principles

Research on the morphological processing of words has focused on two approaches, rule-based systems (e.g. Pentheroudakis and Vanderwende, 1993) and connectionist techniques (Rumelhart and McClelland, 1986; Gasser, 1994). The aim of the AMP system (whose structure is outlined in Fig. 1) is to perform the segmentation of a given set of words (hereafter called a wordset) into stems and endings in an automated manner. To achieve that, a rule-based iterative masking-and-matching approach is used, which relies on matching parts of different patterns while ignoring the remainders of these patterns. For example, if two patterns x_1x_2 and y_1y_2 are to be compared, the technique might focus on the possible similarity between x_1 and y_1 (attempting to match these parts) while ignoring x_2 and y_2 (temporarily masking off those parts). According to that approach, the stemming starts with an initial set of valid stems and/or endings. A fundamental assumption is that each word consists only of a stem part and an ending part, the word being formed by concatenating these parts:

$$<\text{word}> = <\text{stem}> + <\text{ending}>. \qquad (1)$$

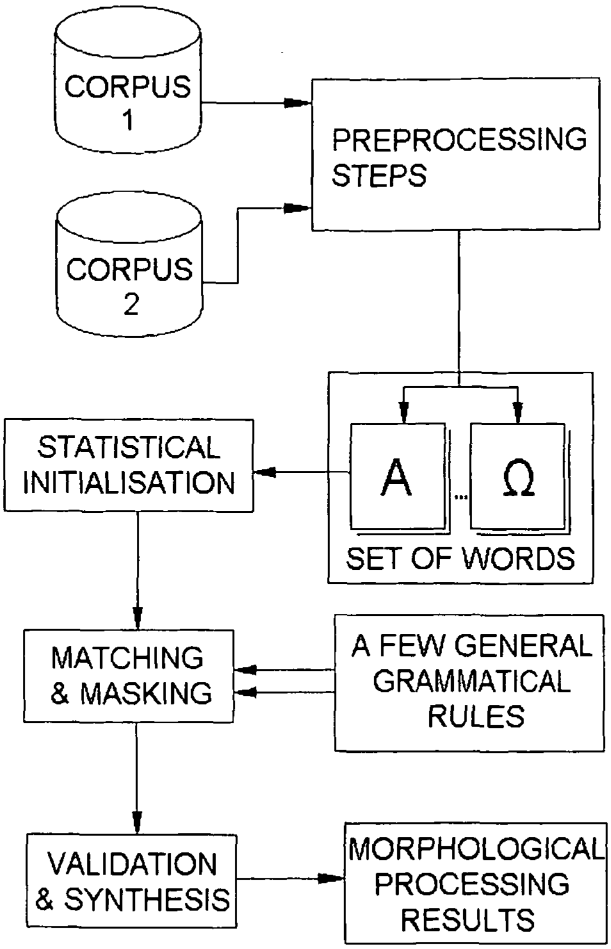


Fig. 1 Block diagram of the AMP system architecture.

The '+' sign in (1) is used to indicate a direct concatenation, without any alteration to either the constituent stem or the ending. When concatenating a stem and an ending, it is possible that the final letter(s) of the stem and the first letter(s) of the ending may interact, in which case the stem and ending will be modified within the word. This type of interaction occurs relatively rarely and therefore, for the sake of simplicity, the approximation expressed by (1) is assumed to hold, allowing the simple implementation of the system. The experimental results obtained by ignoring such interactions (which are reported in Section 6) have confirmed that this simplification leads to a very accurate segmentation (the stemming accuracy being approximately 95 per cent). None the less, as a result of the system's modular structure, the rules of interaction between characters may be introduced in a straightforward manner, further improving the system accuracy (such extensions are studied in more detail in Section 9.2). Currently, following the simplification expressed by (1) and provided that a known stem matches the first characters of the word, the remaining part of the word may be considered as a probable ending. This is expressed by the following operation:

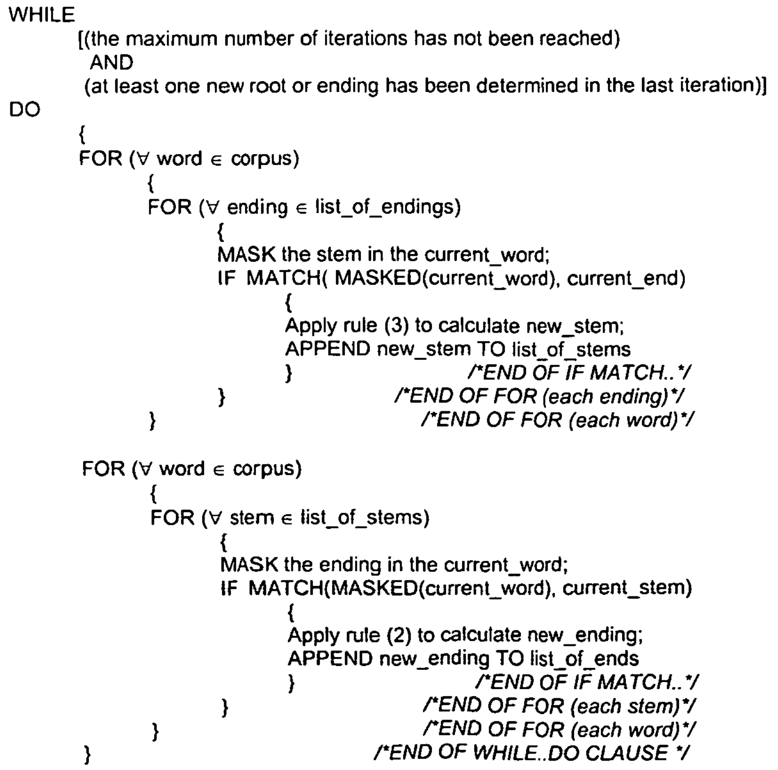
$$<\text{ending}> = <\text{word}> - <\text{stem}>. \qquad (2)$$

The string-difference operator '$-$' used in (2) is applied to the leftmost characters of the word. Following the application of (2) and provided that the corresponding stem is known, one possible ending is calculated. Alternatively, if a known ending has a length of *n* characters and matches the *n* final characters of the word, then a possible stem may be calculated as

$$<\text{stem}> = <\text{word}> \div <\text{ending}> \qquad (3)$$

where '$\div$' indicates a string difference operator applied to the rightmost characters of the word. Rules (2) and (3) are combined to form the iterative scheme of matching-and-masking operations, which is described in Fig. 2 using pseudocode.

During each matching-and-masking iteration, the number of possible segmentations of each word into stem-and-ending pairs rises monotonically. In the Greek language, the morphological phenomenon of derivation, together with inflection (Ralli, 1986), leads to the formation of different wordforms from the same stem. Hence, a word may contain several valid stems of the Greek language, as shown in Table 1. In such a



```
WHILE
        [(the maximum number of iterations has not been reached)
         AND
         (at least one new root or ending has been determined in the last iteration)]
DO
        {
        FOR ($\forall$ word $\in$ corpus)
                {
                FOR ($\forall$ ending $\in$ list_of_endings)
                        {
                        MASK the stem in the current_word;
                        IF MATCH( MASKED(current_word), current_end)
                                {
                                Apply rule (3) to calculate new_stem;
                                APPEND new_stem TO list_of_stems
                                }               /*END OF IF MATCH..*/
                        }               /*END OF FOR (each ending)*/
                }                       /*END OF FOR (each word)*/

        FOR ($\forall$ word $\in$ corpus)
                {
                FOR ($\forall$ stem $\in$ list_of_stems)
                        {
                        MASK the ending in the current_word;
                        IF MATCH(MASKED(current_word), current_stem)
                                {
                                Apply rule (2) to calculate new_ending;
                                APPEND new_ending TO list_of_ends
                                }               /*END OF IF MATCH..*/
                        }                       /*END OF FOR (each stem)*/
                }                               /*END OF FOR (each word)*/
        }                               /*END OF WHILE..DO CLAUSE */
```

Fig. 2 Description of the iterative matching-and-masking process in pseudocode.

Table 1 Examples of words from a corpus (wordset1) with multiple possible stems, all of which are valid stems of the Greek language

Original word form	Stem	Example	Meaning
Δημιουϱγιϰότητα	*δημιουϱγιϰότητ-*	*δημιουϱγιϰότητ+α*	*Creativeness*
	δημιουϱγιϰ-	*δημιουϱγιϰ+ός*	*Creative*
	δημιουϱγ-	*δημιουϱγ+ός*	*Creator*
	δημι-	*δημι+ος*	*Executioner*
	δημ-	*δήμ-ος*	*Municipality*
Πολιτισμιϰός	*πολιτισμιϰ-*	*πολιτισμιϰ+ός*	*Cultural*
	πολιτισμ-	*πολιτισμ+ός*	*Culture/civilization*
	πολιτ-	*πολίτ+ης*	*Citizen*
	πολ-	*πόλ+η*	*City*
Οϱγανωτιϰός	*οϱγανωτιϰ-*	*οϱγανωτιϰ+ός*	*related to organizing*
	οϱγανωτ-	*οϱγανωτ+ής*	*Organizer*
	οϱγαν-	*όϱγαν+ο*	*Instrument*
	οϱγ-	*οϱγ+ή*	*Rage*
	οϱ-	*όϱ+ος*	*Term/mountain*

The examples listed in the third column illustrate a valid wordform originating from the corresponding stem. The variation of the meaning of the different stems (and the corresponding wordforms) is notable.

case, the system will determine several possible segmentations for each word and therefore there is a need for a criterion to establish the most plausible morphological composition, as described in Section 3.

An important property of the AMP system is its modular design. As noted above, the *a priori* knowledge provided to the system (for example, a set of stems or endings) may differ for each language sample. At the same time, although the majority of grammatical constraints remain unchanged for different LESs, there needs to be a provision to easily replace such constraints as required, depending on the LES period for which the system is optimized. This is ensured by providing the AMP with a number of interchangeable modules, which may be selected according to the current LES requirements, whereas the core of the system remains the same. The interchangeable modules are of a declarative nature and may incorporate *a priori* knowledge. Although the principles of the AMP can be applied to different LESs of the Greek language, the current work has focused on the modern Greek language (Triantafillides, 1941) and consequently certain implementation details described hereafter refer to modern Greek.

The AMP principles are not limited to the Greek language, or to languages with a relatively rich morphology. For example, it can be seen that the AMP matching-and-masking principle may be applied to the English language. Let us suppose that the AMP is presented with the following list of verb forms:

{[I] limit, [he] limits, [they] limited}.

Then, by applying rules (2) and (3) it generates the following list of segmentations:

{[I] limit + _, [he] limit+s, [they] limit+ed}

where the symbol '_' signifies a zero-length string.

On the other hand, let us suppose that the AMP is presented with the following list of noun forms:

{dog, dogs, cat, cats}.

Then the AMP is able to perform the following segmentation:

{dog + _, dog + s, cat + _, cat + s}.

It can be seen that in both the cases of verbs and nouns for the English language, zero-length endings may occur. These simple examples indicate that the AMP system is able to process other languages apart from Greek and therefore the technique is to a large extent language independent.

3 Synthesis of Stemming Results

To describe the AMP operation, the following representations are used: an n-letter word is represented as $<l_1l_2 \ldots l_n>$, an m-letter ending is represented as $<e_1e_2 \ldots e_m>$, and a p-letter stem is represented as $<s_1s_2 \ldots s_p>$. As the AMP is intended to generate for each word a stem-and-ending pair, it needs to determine, from all combinations of possible stems and endings, the pair most likely to be correct. To achieve this, the synthesis process described in Fig. 3 is employed, ordering the possible solutions (combinations) according to a classification criterion. The following classification criteria have been evaluated.

(1) *The maximum frequency of the ending.* This criterion selects the solution for which the ending has the highest frequency of occurrence within the given wordset. This criterion is more effective provided that

```
FOR (∀ word ∈ corpus)
     {
     FOR (∀ stem ∈ list_of_stems)
          {
          MASK the ending in the current_word
          IF  MATCH(MASKED(current_word), current_stem)
               {
               FOR (∀ ending ∈ list_of_endings)
                    {
                    Apply rule (1) to GENERATE possible_word;
                    IF MATCH(current_word, possible_word)
                         {
                         APPEND <current_stem,current_ending>
                           TO list_of_segmentations;
                         ORDER list_of_segmentations according to
                           criterion;
                         }              /*END OF IF MATCH(current*/
                    }                   /*END OF FOR (∀ ending*/
               }                 /*END OF IF MATCH (masked current..*/
          }                             /*END OF FOR (∀stem)*/
     }                                  /*END OF FOR (∀ word)*/
```

Fig. 3 Description of the synthesis process in pseudocode.

endings (and their relative frequencies) can be determined more precisely than stems, which holds in the majority of cases:

$$\text{criterion 1: } \max_{\text{all solutions}} \{\mathit{frequency}(<e_1 e_2 \text{☞} e_m>)\}. \tag{4}$$

(2) *The maximum frequency of the stem.* This criterion selects the solution for which the stem has the highest frequency of occurrence. This criterion would be more effective if stems could be determined more precisely than endings:

$$\text{criterion 2: } \max_{\text{all solutions}} \{\mathit{frequency}(<s_1 s_2 \text{☞} s_m>)\}. \tag{5}$$

(3) *The maximum combined frequency of the stems and endings.* This criterion selects the solution(s) where the ending and stem collectively possess the highest frequency of occurrence. The combination of the frequencies of stem and ending parts may be implemented by direct summation of the two frequencies or by a weighted summation, which emphasizes the frequency of the part that is calculated more accurately:

$$\text{criterion 3: } \max_{\text{all solutions}} \{f(\mathit{frequency}(<s_1 s_2 \text{☞} s_m>), \text{frequency}(<e_1 e_2 \text{☞} e_m>)\}. \tag{6}$$

(4) *The minmax frequency of the stem and ending parts.* This criterion selects the solution for which the minimum of the frequencies of occurrence of the stem and the ending is maximized. This criterion is a special case of criterion 3:

$$\text{criterion 4: } \max_{\text{all solutions}} \{\min(\mathit{frequency}(<s_1 s_2 \text{☞} s_m>),\text{frequency}(<e_1 e_2 \text{☞} e_m>)\}. \tag{7}$$

(5) *The minimum-length ending.* This criterion selects the solution that contains the shortest possible ending:

$$\text{criterion 5: } \{ \min_{\text{all solutions}} (\text{n}), \text{where n} = \text{length}\,(<e_1 e_2 \text{☞} e_m>)\}. \tag{8}$$

Of course, more than one criterion may be combined and applied at the same time, to improve the system performance. Indeed, both criteria 3 and 4 are weighted combinations of criteria 1 and 2. The best results have been obtained by applying criteria 4 and 5, and therefore the experiments presented here will focus on these two criteria.

4 The Pre-processing Stage

The corpus that is to be processed by the AMP is presented to it as a sequence of words. To that end, the corpus is pre-processed to remove information regarding the text formatting as well as punctuation marks. Additionally, the following filtering operations are performed to remove any invalid words.

(1) Words containing characters other than the letters of the Greek alphabet are removed, as they are not valid words of the Greek language:

$$\text{if } \exists l_i, (1 \leqslant \text{i} \leqslant \text{n}){:}\ l_i \notin \{\alpha, \beta, \text{☞} \omega, \text{A}, \text{B}, \text{☞} \Omega\}, \text{ then remove } <l_l l_2 \text{☞} l_n>. \tag{9}$$

(2) Invariant (i.e. not conjugated) words are removed. For the purposes of categorization, when creating the ILSP morphological lexicon (Gavrilidou, 1996), invariant words have been organized into the following categories: nouns and adjectives that are conjugated but whose formation is irregular; foreign-origin nouns and adjectives; conjunctions; exclamatives; numerals; invariant nouns and adjectives; adverbs; particles; and pronouns. The majority of these words are derived from the stems of conjugated words (for example, the adverb «τελεί-ως»- meaning *entirely*, has a common stem with the adjective «τέλει-ος»- meaning *perfect*). The segmentation of these words by the AMP is allowed to take place, to reveal such derivations (the issue of derivation is further discussed in Section 9.1). The list of invariant words provided to the AMP consists of conjunctions, particles, prepositions, and pronouns. Within these categories, only the most frequently occurring members not derived from conjugated words are retained, resulting in a considerably reduced list (with 182 elements as opposed to the 1912 invariant words contained in the ILSP lexicon).

(3) For all words that are not filtered out in steps (1) or (2), any upper-case letters are replaced by the corresponding lower-case ones, and characters with stress are replaced by the same characters without stress (for example, *«Λάθος»* becomes *«λαθος»*). The replacement of upper-case letters may lead to the segmentation of some invariant words (for example, names of foreign towns such as *«Τόκιο»* - *Tokyo*). Removing the stress from words may cause some ambiguities, although the stress may be retrieved from the input stream at a later stage to resolve any ambiguities if required. These approximations do not affect the system performance, as will be shown by the experimental results. It should be noted that in current experiments, a different approach to the stress information is followed. Hence, although the initial AMP stemming is still based on separating the stress from the word, the stress information is stored and is employed when required, to provide additional information for disambiguation purposes as well as to provide an increased accuracy in morphological processing. Nevertheless, in the following experimental results, where the stress information is not employed, the stemming accuracy still remains high.

(4) Any word that occurs in the text more than once is included in the wordset only in the place of its first appearance.

Hence, following this pre-processing stage, the wordset comprises words each of which consists solely of Greek characters. Words with a higher frequency of occurrence can be expected to be nearer the beginning of the wordset, less frequent ones being positioned as a rule nearer the end of the wordset. This fact necessitates the study of the AMP performance throughout the wordset.

It should be noted that even after this pre-processing, the original corpus remains available. Thus the information contained in the environment of the word within the original text may be used to retrieve information regarding the word, ranging from part-of-speech (the simplest case) to gender or number (using more detailed information). Thus, in cases where more than one possible segmentation into stem and ending is located for a given word, such information may be used to improve the system accuracy. In the remainder of this paper, information obtained via the environment will not be employed. Indeed, as will be shown by the experimental results, even without using this information the system stemming accuracy remains sufficiently high. Therefore, the introduction of environment information at a later stage may only improve the AMP accuracy over the results presented in the following experiments.

5 Implementation Details

5.1 System initialization

To generate more accurate morphological results, the AMP starts processing the wordset with an initial set of known stems and/or endings (provided as *a priori* knowledge). This set is progressively augmented by the solutions determined via the masking-and-matching iterations. It has been found that, in the absence of any other information, three endings form a sufficient initialization set from which the AMP is able to successfully process the wordset.

An initializing set of endings and/or stems may be calculated from the wordset by statistical methods, based on the frequency of occurrence of *n*-grams (combinations of *n* consecutive characters). Digrams and trigrams have been used to generate a subset of French words from a lexicon, to provide a training set for pattern recognition studies (Barriere and Plamondon, 1998). Greffenstette (1995) has proposed studying the last two letters of each word (hereafter this pair will be referred to as the end-digram) to determine the language in which a specific text has been written, indicating that the end-digram is a fundamental characteristic of each language.

Although valid endings in the modern Greek language can exceed seven characters in very rare cases, the majority of endings in the Greek language consist of between one and three characters (especially in nouns and adjectives, verbs tending to have longer endings). Single-character endings are considered to be too small to provide reliable results. On the other hand, end-trigrams are less frequent and the computational effort required to determine them is larger. For a language with N characters the number of possible digrams is N^2, whereas the number of possible trigrams rises to N^3, necessitating a considerably larger corpus to calculate accurately the frequency of occurrence of end-trigrams. Consequently, end-digrams have been selected as one possible source of initialization data.

During initialization, all the wordset members are read and the frequencies of occurrence of end-digrams are calculated. The end-digrams with the highest frequencies are then selected as probable endings. Grammatical rules might be used to reject invalid endings (for example, as endings in modern Greek contain at least one vowel, end-digrams consisting only of consonants cannot be valid endings). However, experimentation has shown that, as a rule, for sets with a relatively large number of words (more than 1000 words), the most frequent end-digrams are valid endings (typical results are provided in Section 6). Thus, the three most frequent end-digrams are used to initialize the system, supplementing any other *a priori* information.

Of course, the *a priori* knowledge provided may vary. Indeed, it would be possible to provide as *a priori* knowledge all the possible endings for a given LES. Then, the masking-and-matching algorithm could generate in a single pass all the stems in the new corpus, processing unknown word-forms and eliminating the need for the iterative process. On the other hand, the AMP is able to operate successfully even with a very limited amount of *a priori* knowledge. In that case, the AMP stems the words of a corpus with a high degree of accuracy, using the masking-and-matching iterations to progressively extract knowledge from the corpus.

5.2 Constraints inspired from the Greek grammar

Experimentation has indicated that the AMP generates a relatively large number of possible solutions via its matching-and-masking technique (typically, around 17,000 stems and 8000 endings for a given wordset of 10,000 non-replicated wordforms). For each combination of stem and ending, a metric needs to be calculated (according to the corresponding criterion) that represents the likelihood of the given wordform originating from this stem-and-ending pair. The larger the number of candidate stems and endings becomes, the more computations of the likelihood metric need to be performed and the higher the computational load becomes on the simulation environment, in terms of both processing power and storage capacity. Hence, by removing stem-and-ending pairs that do not lead to valid solutions, the processing of larger wordsets may be performed in a significantly reduced time.

To achieve that, a handful of rules are provided regarding valid endings or stems in the Greek language. This is in contrast to the approach by Kay and Roescheisen (1993) in discovering morphological relations, where no such rules are used. However, Kay and Roescheisen propose a text-translation aligning system that is applicable to several languages and therefore they do not focus on optimizing the morphological processing for one specific language. The incorporation of the following constraints regarding the highly inflectional modern Greek language leads to a considerable improvement in the AMP performance.

(1) Endings need to contain at least one vowel. This rule has been found to eliminate a large number of false solutions:

$$\text{if } \forall e_i,\ (1 \leq i \leq m)\text{: } e_i \in \{\text{Greek consonants}\}\text{, then reject } <e_1e_2\ldots e_m>. \quad (10)$$

(2) If a diphthong[1] is contained in a word, the two vowels of that diphthong may be situated either both in the constituent stem or both in the constituent ending, as, according to the Greek grammar (Triantafillidis, 1941), diphthongs may not be separated:

$$\text{if } (s_p,e_1),\ \epsilon\{\text{Greek diphthongs}\}\text{, then reject the pair } (<s_1s_2\text{☞}s_p>,<e_1e_2\text{☞}e_m>). \quad (11)$$

[1] The following vowel combinations are Greek diphthongs (Triantafillidis, 1941): *αι, αυ, ει, ευ, ηυ, οι, ου*.

(3) The length of a stem must be at least three characters. This constraint is imposed by the implementation environment rather than the language grammar:

$$\text{if } p < 3, \text{ then reject stem } <s_1 s_2\text{☞}s_p>. \qquad (12)$$

To illustrate the purpose of this constraint, let us suppose that a one-letter stem $<s_1>$ has been determined at some point of the matching-and-masking application. If constraint (12) is not applied, in the next matching-and-masking step, for all words starting with character s_1, new (but highly improbable) endings would be determined, generating an excessive number of solutions.

The constraint expressed by (12) prevents the correct segmentation of short-stem words (words with a stem of only one or two characters). Such words are very rare and therefore do not affect the AMP accuracy, as will be shown by the experimental results. Still, the stemming of short-stem words may be achieved during the synthesis phase on the basis of the valid endings, marginally increasing the segmentation accuracy.

(4) The length of an ending is constrained to a maximum of seven characters:

$$\text{if } m < 7, \text{ then ignore ending } <e_1 e_2\text{☞}e_m>. \qquad (13)$$

This constraint complements (12) and prevents the generation of an excessive amount of endings, especially for long words. Constraint (13) is valid, with possibly very few exceptions in the Greek language.

(5) The initial letter of any valid ending needs to be a vowel:

$$\text{if } e_1 \notin \{"\alpha", "\varepsilon", "\eta", "\iota", "o", "\upsilon", "\omega"\}, \text{ then ignore ending } <e_1 e_2\text{☞}e_m>. \qquad (14)$$

This constraint has been introduced on the basis of experimental results, as it has been observed that the vast majority of endings start with a vowel. Formula (14) drastically reduces the number of incorrect stem-ending segmentations and contributes to the improvement of the AMP accuracy.

6 Experimental Results

6.1 Characteristics of wordset sources

In this section, the results obtained for two sets of texts are described. The first set (hereafter denoted as wordset1) consists of excerpts from a novel, whereas the second set (denoted as wordset2) is a collation of newspaper articles with subjects ranging from politics to sport and science, published in a daily newspaper on 6 and 7 May 1996. The different origins of the two texts allow the evaluation of the AMP system in two fundamentally different environments. Text obtained from a novel can be expected to have a well-defined theme, to be of a high linguistic quality and without spelling mistakes. Additionally, as the novel is created by a single author, it will use a consistent version of the Greek language and most probably a large proportion of wordforms will be related to each other. In contrast, the collection of newspaper articles is selected on the basis of publication date and therefore the topics can be expected to vary considerably. As manuscripts by several authors are included, a single standard of language does not exist (even different LESs may coexist) and the vocabulary employed is less uniform. The likelihood of spelling errors also increases because of the limited time available for the creation, editing, and publication of each article.

To investigate the accuracy of the AMP results, two versions of the morphological lexicon developed in the ILSP are employed (Gavrilidou, 1996). The first one (denoted as LEX_std) contains approximately 50,000 lemmas of the Greek language and 92,367 allomorphs in total. The second version (denoted as LEX_ext) is an extended version of LEX_std, which contains more than 63,000 lemmas and 107,026 allomorphs in total.

6.2 Performance of the frequency-based initialization

When wordset1 is pre-processed, 48,317 words are read to generate a list of 10,000 non-repeated words. These result in the generation of 183 end-digrams. Similarly, when wordset2 is pre-processed, 33,163 words are read to generate a 7853-word list, which contains 187 end-digrams.

The ten most frequent digrams for each corpus are shown in Table 2, with their respective frequencies of occurrence. It is worth noting that out of the ten most frequent end-digrams of wordset1, six ("*-ες*", "*-ια*", "*-ει*", "*-ος*", "*-ας*", "*-ου*") correspond to valid ends. Of the remaining four end-digrams, three are the least frequent. The sole exception is end-digram "*-νε*", which is part of the valid verb endings "*-ανε*", "*-ουνε*" (third person, plural) and is justifiably very frequent in the text-type of novels, which are characterized by a narrative style. However, provided that a sufficient number of initializing end-digrams correspond to valid endings of the Greek language, the system will process the wordset successfully even if one initializing end-digram is not valid. Out of the ten most frequent end-digrams of wordset2, two are not valid ends (end-digrams "*-υν*", "*-ση*"), both having relatively low rankings. Although the

Table 2 End-digram occurrences for the two wordsets used experimentally; the frequency of occurrence is expressed as a percentage of all end-digram occurrences

	Wordset1		Wordset2	
Rank	End-digram	Frequency (%)	End-digram	Frequency (%)
1	*-ες*	6.60	*-ει*	6.71
2	*-ια*	6.18	*-ης*	5.40
3	*-νε*	6.17	*-ων*	5.38
4	*-ει*	6.16	*-ος*	4.92
5	*-ος*	4.13	*-ου*	4.21
6	*-ας*	3.66	*-ια*	3.83
7	*-ου*	3.23	*-ας*	3.53
8	*-σε*	2.68	*-ην*	3.33
9	*-να*	2.67	*-αν*	2.91
10	*-τα*	2.51	*-ση*	2.77

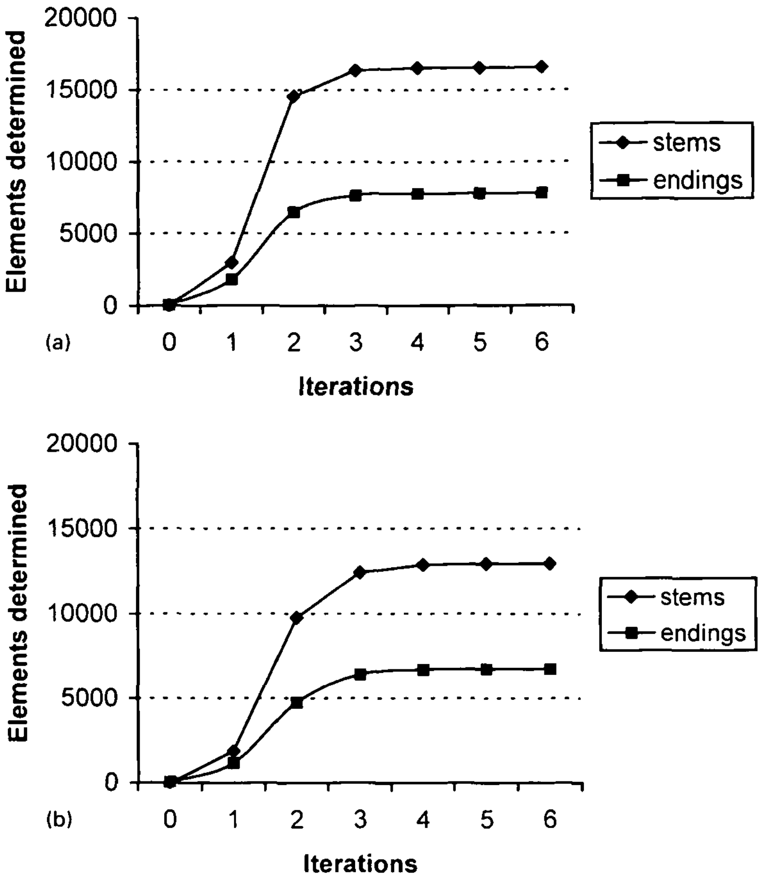


Fig. 4 Stems and endings that have been determined after each masking-and-matching iteration, for wordset1 (a) and wordset2 (b).

three most frequent end-digrams differ for the two wordsets, five valid digrams out of the ten most frequent digrams coincide, indicating the consistency of the proposed frequency-based initialization.

6.3 Results of the AMP System

The AMP system is simulated on one processor of a SUN 450 workstation, the morphological processing of each corpus needing in total between 12 and 20 min. Approximately 20 per cent of the execution time is required for the matching-and-masking iterations and the remainder is required for the synthesis step, which is more computationally intensive, in terms of both processing time and memory requirements.

The calculation of solutions throughout the iterations is summarized in Fig. 4. In each simulation, the AMP settles after, at most, six iterations of the masking-and-matching operation, with most stems and endings being discovered in the second and the third iteration (when more than 70 per cent and 95 per cent of the total elements have been discovered, respectively). Following the completion of the masking-and-matching iterations, the morphological segmentations found for each word are ordered via the ranking criterion. These results are then compared with

Table 3 Comparison between the AMP results and the ILSP lexicon, for various criteria with no *a priori* knowledge; in each case, the distribution of solutions is expressed as a percentage of the total number of words

	First solution	Second solution	Third solution	Fourth solution	Fifth+ solution	Not segmented
Wordset1: minmax frequency criterion	48.7	26.3	13.1	4.5	1.2	6.2
Wordset1: shortest-ending criterion	80.4	5.7	6.6	1.9	1.4	3.8
Wordset2: minmax frequency criterion	46.5	24.4	18.2	7.5	1.2	2.2
Wordset2: shortest-ending criterion	84.6	2.7	7.7	2.5	0.3	2.2

the contents of the standard ILSP morphological lexicon (LEX_std). Approximately 20 per cent and 24 per cent of the wordforms in wordset1 and wordset2, respectively, do not correspond to any entry in the LEX_std lexicon. In Table 3, the AMP results are presented for the first 500 words of wordset1 and wordset2 using (1) the minmax frequency of occurrence criterion and (2) the shortest-ending criterion. In this table, the '*k*th solution' entry (where $k = 1, 2, 3, 4$) indicates the percentage of words for which the *k*th ranked solution of the AMP corresponds to the correct segmentation, as determined by the ILSP lexicon. Cases where the correct segmentation is generated by the AMP but ranked as fifth or lower are denoted as 'fifth+ solution'. The cases for which the AMP fails to generate the correct solution are annotated as 'not segmented'.

As can be seen from Table 3, the AMP displays a similar performance for both texts. Using the minmax frequency criterion, almost 50 per cent of the words are correctly segmented by the top-ranked AMP solution. A substantial fraction of the words (around 25 per cent) are segmented via the second-highest solution, approximately 15 per cent via the third solution, 5 per cent via the fourth solution, and only very small amounts via lower-ranking solutions. The percentage of words that the AMP fails to segment ranges from 2 to 6 per cent, and is thus acceptably low, whereas most words (around 85–90 per cent) are correctly segmented via one of the three top-scoring AMP solutions.

By using the shortest-ending criterion for the ordering of possible solutions, a considerable improvement is achieved. According to Table 3, approximately 80–85 per cent of the words are segmented correctly by the highest-scoring AMP solution, around 3–6 per cent by the second-highest solution, and 7–8 per cent by the third-highest solution. Solutions ranked fourth or lower are very rarely the correct ones, in both wordsets representing no more than 3.5 per cent of all segmented words.

The shortest-ending criterion proves to be the most efficient one, this being in agreement with Kay and Roescheisen (1993). However, using this criterion, the correct segmentation is not chosen when a sub-optimal one involves a shorter ending. Hence, a significant improvement may be achieved if at least some of the longer endings are determined and

Table 4 Comparison between the AMP results and the ILSP lexicon, for the shortest-ending criterion with and without *a priori* knowledge; in each case, the distribution of solutions is expressed as a percentage of the total number of words

	First solution	Second solution	Third solution	Fourth solution	Fifth+ solution	Not segmented.
Wordset1: no *a priori* knowledge	80.4	5.7	6.6	1.9	1.4	3.8
Text 1: with *a priori* knowledge	93.6	2.6	0.0	0.0	0.0	3.8
Wordset2: no *a priori* knowledge	84.6	2.7	7.7	2.5	0.3	2.2
Text 2: with *a priori* knowledge	95.8	1.7	0.3	0.0	0.0	2.2

provided to the system as *a priori* knowledge. This knowledge can be obtained easily for a given LES from any handbook of the Greek grammar (e.g. Triantafillidis, 1941). The results obtained for the two wordsets using a subset of forty 'long' endings are shown in Table 4. Compared with the original results without any *a priori* knowledge, the accuracy of the highest-ranked solution is approximately 95 per cent. At the same time, the accuracy of lower-ranked solutions is reduced drastically, third- or lower-ranked solutions having a zero probability of use. This indicates that by introducing *a priori* knowledge, if the correct segmentation can be determined, it will be either the first-ranked solution or (rarely) the second-ranked solution.

6.4 System behaviour through time

The AMP results are expected to vary as the wordset is being processed, as the more frequent words are expected to be nearer the beginning of the wordset. The AMP performance for the shortest-ending criterion with *a priori* knowledge is shown in Fig. 5a and b, for wordset1 and wordset2, respectively. Although the actual distribution of the solutions varies, the AMP behaviour remains the same, with the top-scoring solution being most frequently the correct segmentation (its frequency ranging from 90 to 97 per cent). The second-ranked solution has a frequency of between 1 and 3 per cent, whereas the lesser-ranked solutions have a frequency very close to zero.

Diagrams illustrating the AMP performance when no *a priori* knowledge is provided to the system and the shortest-ending criterion is employed are shown in Fig. 5c and d, for wordset1 and wordset2, respectively. In this case, the highest-ranked solution has a frequency of between 76 and 88 per cent. Hence, the frequency margin increases, although the highest-ranked solution remains in most cases the correct segmentation. The second- and third-ranked solutions have broadly comparable frequencies ranging from 2.5 to 10 per cent, whereas the fourth- and fifth+-ranked solutions have a lower (but non-zero) fre-

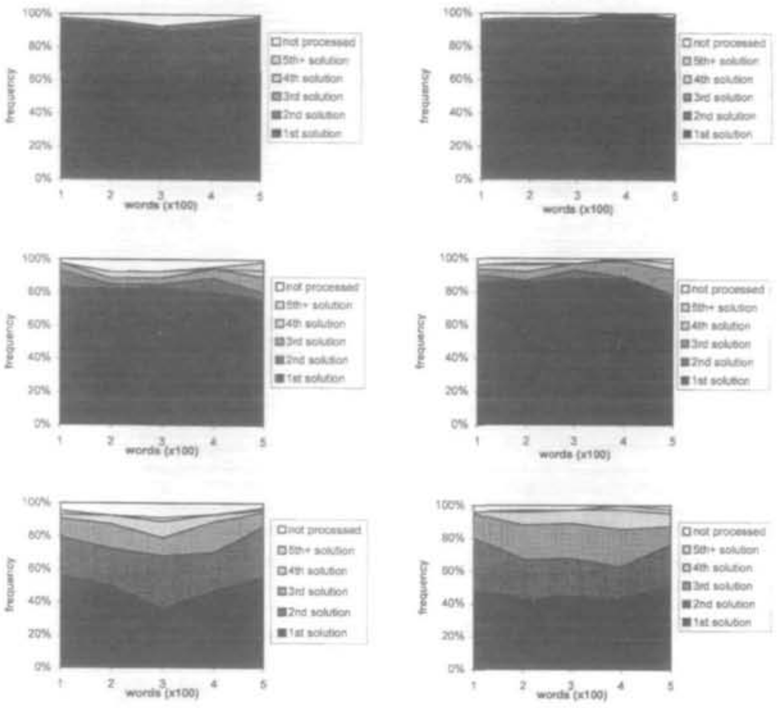


Fig. 5 Evolution of frequencies of the ranked solutions during the stemming of a wordset. (a) and (b) refer to the minimum-length ending criterion with *a priori* knowledge for wordset1 (a) or wordset2 (b). (c) and (d) refer to the minimum-length ending criterion without *a priori* knowledge for wordset1 (c) or wordset2 (d). (e) and (f) refer to the minmax frequency criterion without *a priori* knowledge for wordset1 (e) or wordset2 (f).

quency. Although these frequencies vary considerably, the general characteristics of the solution distribution remain unchanged while each wordset is processed.

The diagrams obtained for wordset1 and wordset2 when no *a priori* knowledge is provided to the system and the minmax frequency of occurrence criterion is used are shown in Fig. 5e and f, respectively. This ranking criterion, in the absence of *a priori* knowledge, is found to generate the results with the highest variability during processing.

7 Extended Stemming Experiments

As noted at the end of Section 6.3, by combining the shortest-ending criterion with a small amount of *a priori* knowledge, approximately 95 per cent of all words are segmented correctly by the first solution of the AMP. To confirm the validity of these observations, experiments involving larger corpora have been carried out.

To investigate the effect of *a priori* knowledge on the AMP accuracy, a series of experiments have been performed, where the set of *a priori* endings is modulated from 0 per cent to 100 per cent, all other parameters remaining unchanged. The corpus consists of novels, and has a total size of 485,955 words, resulting in a list of 43,150 different wordforms. The results summarized in Table 5 indicate how the addition of *a priori* knowledge improves the AMP accuracy. This is attributable to the highly inflectional nature of modern Greek in comparison with other languages.

Table 5 Distribution of AMP solutions in comparison with the LEX_std lexicon, for a corpus of 490,000 words, using the shortest-ending criterion (LEX_std) (values are given as percentages)

A priori information	First solution	Second solution	Third solution	Fourth solution	Not segmented
0	79.9	4.3	11.1	3.1	1.6
25	82.7	2.4	10.9	2.8	1.2
50	85.9	2.5	10.2	0.4	1.0
75	91.1	2.4	5.1	0.4	1.0
100	93.6	4.2	0.9	0.3	1.0

An important point concerns the effect of the corpus size on the AMP accuracy. Therefore, further experiments have been carried out to determine whether the system accuracy will be affected when using a substantially larger corpus. For this, a corpus is formed containing excerpts from a total of twenty-eight literary novels and scientific publications. This combination of different genres is essential to achieve a corpus of 17 Mbytes, containing a total of 2,530,296 words. Following the AMP pre-processing stage, this corpus results in a set of 106,093 different wordforms of the Greek language. Out of these, 31,007 wordforms do not have an entry in the ILSP lexicon. The AMP generates a total of 265,038 solutions, using 75,518 roots and 26,577 ends. The results obtained by the AMP system are displayed in Table 6. By comparing these with Table 5, it is evident that the AMP retains its ability to generate highly accurate results even when operating on large corpora.

8 Automated Extraction of Stemming Information for Corpora Possessing Terminology Wealth

In this section, the AMP is employed to generate specialized morphological lexica using a corpus of documents (Tambouratzis and Carayannis, 1999). The construction of a lexicon by hand is a particularly arduous task, as it requires the creation of a list containing all possible wordforms in the respective language. The completion of this list requires several person-months, and it remains virtually impossible to provide a complete coverage of the language, because of its constant evolution. Therefore, the majority of morphological lexica used in the framework of the various applications are intended to cover the most frequent words of the language. This coverage becomes substantially lower in cases where the source documents refer to a specialized area such as a particular scientific field.

The AMP is presented with a corpus of documents and generates the corresponding morphological lexicon in an automated manner, by gradually segmenting the words of this corpus into roots and endings via successive matching-and-masking iterations. Of course, the degree of coverage provided by the lexicon as well as the lexicon completeness depends on the corpus. Although the resulting lexicon will probably require some correction by hand, most of the correct entries are generated automatically, considerably reducing the human workload.

To study the efficiency of this process, three corpora are used, denoted as C1, C2 and C3. Corpus C1 consists of novels and thus is expected to employ an LES that should be covered satisfactorily by general-purpose morphological lexica. Corpus C2 consists of three textbook excerpts on aerospace engineering, and C3 consists of three textbook excerpts on biology. C2 contains a total of 245,006 words, and C3 a total of 282,902 words. Corpus C1 consists of 199,494 words, to provide an amount of wordforms similar to that of C2 and C3. The number of different wordforms is 20,000 for C1, 13,905 for C2, and 20,804 for C3.

Both the standard and the extended versions of the ILSP morphological lexicon are employed to evaluate the accuracy of the proposed method. As shown in Table 7, the amount of wordforms unknown to general-purpose lexica is considerably higher in corpora consisting of term-intensive documents (in particular, corpus C3). The wordforms that the AMP fails to separate into stem and ending are approximately 1.5 per cent in the case of C2 and 1 per cent in the case of C3. These amounts are similar to that of corpus C1, indicating that the AMP exhibits a consistency for different corpora and provides a considerably higher coverage of corpus wordforms than the two general-purpose lexica.

The accuracy of the solutions generated by the AMP in comparison with the contents of the two ILSP morphological lexica is depicted in Table 8, where the corresponding fractions refer to the wordforms that are contained in each morphological lexicon. For approximately 94 per cent of the wordforms, the entry of the morphological lexicon coincides with the highest-scoring AMP segmentation. For approximately 4–5 per cent of the wordforms, the segmentation proposed by the morphological lexicon coincides with the second-highest solution of the AMP. Cases

Table 6 AMP results for a corpus of 2.5 million words, using the shortest-ending criterion and the LEX_std morphological lexicon

First solution	Second solution	Third solution	Fourth solution	Not segmented
70 720 (93.6%)	4022 (5.4%)	170 (0.2%)	5 (0.0%)	619 (0.8%)

Table 7 Coverage of wordforms from each corpus (values are given as percentages)

	Corpus C1	Corpus C2	Corpus C3
Lexicon LEX_std	89.0	84.2	74.4
Lexicon LEX_ext	90.9	86.3	76.3
AMP system	98.9	98.6	99.3

Table 8 Comparison of results of the AMP system to the contents of the morphological lexica (values are given as percentages)

	First solution	Second solution	Third solution	Fourth solution	Not segmented
C1 & LEX_std	93.7	4.9	0.3	0.0	1.1
C1 & LEX_ext	93.7	4.8	0.4	0.0	1.1
C2 & LEX_std	93.9	4.4	0.1	0.0	1.6
C2 & LEX_ext	93.8	4.3	0.2	0.0	1.7
C3 & LEX_std	94.0	4.9	0.1	0.0	1.0
C3 & LEX_ext	93.9	5.0	0.1	0.0	1.0

where the lexicon entry coincides with the third-highest AMP solution are extremely rare (around 0.1 per cent), and the fourth AMP solution (or any lower-ranked solution) does not coincide with the lexicon entry for any wordform. It is worth noting the stability of the distribution of the correct stemming over the three corpora, when using both lexica.

These results confirm that the automated morphological processor generates the grammatically correct solution for the vast majority of wordforms encountered in the corpora. Thus, the AMP may be employed to accurately and consistently generate morphological data from corpora of term-intensive documents that focus on a particular field.

9 Extending the AMP Structure

9.1 Extending the AMP to describe the derivation phenomenon

The work described in this paper can be extended to increase the AMP functionality. A likely candidate is the use of the matching-and-masking technique to record the compounding and derivation phenomena in the Greek language. To that end, the list of stems generated by the AMP may be used in a subsequent processing step. In that case, the masking and matching rules (2) and (3) need to be extended. As an example, the derivational relationship between the four stems {'*πολιτισμικ*', '*πολιτισμ*', '*πολιτ*', and '*πολ*'} (included in Table 1) can be determined by using the following formula:

$$\text{<inflection>} = \text{<stem2>} - \text{<stem1>}. \qquad (15)$$

Formula (15) may be applied only in the case of two stems, stem1 and stem2, where stem2 contains stem1 and the operator '−' indicates the string difference obtained by subtracting stem1 from the leftmost part of stem2. Hence, by applying rule (15), the derivation of '*πολιτ*' from stem '*πολ*' by adding the string '*ιτ*' is uncovered. The use of prefixes can be formulated in a similar way:

$$\text{<prefix>} = \text{<stem2>} \varnothing \text{<stem1>}. \qquad (16)$$

where the symbol ∅ indicates the string difference between stem2 and stem1 starting from the rightmost part of stem2 (for example, the addition of prefix '*α*' to stem '*πολιτιϰ*' to form the stem '*απολιτιϰ*', as used in the wordform '*απολιτιϰός*' meaning non-political). Similarly, the compounding operation to form the word '*δεϰαεννι+ά*' (nineteen) from the word '*δέϰ+α*' (ten) and '*εννι+ά*' (nine) could be uncovered via an extension of the matching-and-masking operator. In this case, the rule could be generalized to a form similar to

$$<\text{stem3}> = <\text{stem1}> \oplus <\text{stem2}>. \tag{17}$$

It should be noted that in the operations described by (15)–(17), interactions between the strings involved might occur. Therefore, these phenomena and formulae need to be studied in more detail and will form the basis of subsequent research, leading to a knowledge-rich extension of the AMP system with additional functionality. On the other hand, the work presented in this paper indicates how a less complex system provided with only a minimum amount of grammatical knowledge may be used to successfully process corpora and provide the basis for automatically generated morphological information.

9.2 Extending the linguistic information to cover the interaction between characters

The experimental results reported above have indicated that the AMP system can perform stemming with a high degree of accuracy. However, there are certain phenomena (in particular, in verbs) that may not be covered accurately, by ignoring all interactions between the last characters of the stem and the first characters of the ending. To cover such cases, the language-specific knowledge provided *a priori* needs to be increased. For example, the provision of a more comprehensive list of endings can help address this issue. Additionally, rules covering the interaction of letters may need to be provided, for the system to detect the relationship between superficially different stems. However, the most effective method for detecting these interactions is, in our opinion, to provide a set of fuzzy matching rules when comparing stems. These rules coupled with linguistic information regarding the interaction between characters from the stem and ending can provide a more accurate model for stemming in the Greek language.

10 Conclusions

In this paper, a system has been presented that allows the morphological processing of a set of words belonging to the Greek language in an automated manner, based on a statistical initialization followed by a masking-and-matching algorithm. This system accurately segments each word into a stem and an ending using some *a priori* knowledge and thus allows the generation of morphological lexica faster and with considerably reduced human effort, in comparison with hand-made lexica. As the AMP extracts knowledge from the input text, it is able to morphologically process rare or idiomatic wordforms with success. A statistical method for initializing the AMP has also been proposed, based on the frequency of occurrence of the end-digram of each word. This method has been shown to provide an accurate initialization for the AMP system.

During operation, the AMP generates a number of possible solutions. To select the one that represents the correct segmentation, a ranking criterion is employed. In this work, several such criteria have been introduced and evaluated, using the existing ILSP morphological lexicon of the Greek language for comparison purposes. The best results are obtained using a shortest-ending criterion, with a modest amount of *a priori* knowledge, for which the segmentation accuracy is approximately 95 per cent. The AMP is able to segment previously unseen words, providing a much higher coverage than the ILSP lexicon. The proposed AMP system has been applied to term-intensive corpora of texts and has been shown to successfully generate specialized morphological lexica in an autonomous manner, achieving a coverage of approximately 99 per cent of the words from each corpus and consequently resulting in a substantial reduction of the associated workload.

A case study has indicated that the AMP principles can be used for stemming operations in other languages, with considerable success. Finally, a study has been presented of possible extensions to the AMP system. These allow it to cover additional morphological operations such as derivation, indicating that it may form the basis for a general-purpose morphological processing platform.

Acknowledgements

The authors wish to acknowledge the assistance of Dr. Christos Stathis in making available specialized programs to separate the input words into stems and endings according to the entries of the ILSP morphological lexicon for comparison purposes. Additionally, the authors wish to acknowledge the constructive comments of the anonymous referees, which have led to an improved version of the paper.

References

Barriere, C. & Plamondon, R. (1998). Human identification of letters in mixed-script hand-writing: an upper bound on recognition rates. *IEEE Transactions on Systems, Man and Cybernetics—Part B: Cybernetics*, 28(1): 78–82.

Gasser, M. (1994). Modularity in a connectionist model of morphology acquisition. In *Proceedings of the 15th International Conference on Computational Linguistics, 5–9 August 1994, Kyoto, Vol. 1*, pp. 214–20.

Gavrilidou, M. (ed.) (1996). The ILSP morphological lexicon and morpho-syntactic tagger. Internal report. Athens: ILSP (in Greek).

Greffenstette, G. (1995). Comparing two language identification schemes. In Bolasco, S., Lebart, L. and Salem, A., *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95), Rome, December 1995*. Rome: CISU. Available at: http://www.rxrc.xerox.dom/publis/mltt/jadt/jadt.html.

Kay, M. & Roescheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, 19(1): 121–42.

Pentheroudakis, J. & Vanderwende, L. (1993). Automatically identifying morphological relations in machine-readable dictionaries. Technical report MSR-TR-93-06. Redmond, WA: Microsoft Research Advanced Technology Division, Microsoft Corporation.

Ralli, A. (1986). Derivation and inflection. In *Studies in Greek Linguistics—Proceedings of the 7th Annual Meeting of the Department of Linguistics, Faculty of Philosophy, Aristotelian University of Thessaloniki, 12–14 May 1986*, pp. 29–48 (in Greek). Thessaloniki: Kiriakidis.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tense of English verbs. In McClelland, J. L. & Rumelhart, D. E. (eds), *Parallel Distributed Processing, Vol. 2.* Cambridge, MA: MIT Press, pp. 216–71.

Tambouratzis, G. & Carayannis, G. (1999). Automated construction of morphological lexica possessing terminology wealth on the basis of term-intensive documents. In *Proceedings of the Second Conference on Greek Language and Terminology, 21–23 October, Athens*, pp. 149–56 (in Greek). Athens: Hellenic Society for Terminology.

Triantafillidis, M. (1941). *Modern Greek Grammar (Dimotiki)*. Reprint with corrections, 1978. Thessaloniki: Institute of Modern Greek Studies (in Greek).