International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:6, 2007

Greek Compounds: A challenging case for the parsing techniques of PC-KIMMO v.2

Angela Ralli and Eleni Galiotou

***Abstract*—** In this paper we describe the recognition process of Greek compound words using the PC-KIMMO software.

We try to show certain limitations of the system with respect to the principles of compound formation in Greek. Moreover, we discuss the computational processing of phenomena such as stress and syllabification which are indispensable for the analysis of such constructions and we try to propose linguistically-acceptable solutions within the particular system.

***Keywords*—** Morpho-phonological parsing, compound words, two-level morphology, natural language processing.

I. Introduction

This paper deals with a computational analysis of compounds in Modern Greek (hereafter Greek). Greek compounding displays a particular interest on both theoretical and computational grounds since it interacts with derivation, inflection and phonology to a significant extent. To our knowledge, there are no other computational works available dealing with their internal structure.

The computational analysis, which is still at a prototype level, is based on PC-KIMMO version 2, a well-known tool for its world-wide applications to the morphology of several languages. The current state of the implementation handles the recognition of nominal and verbal compounds in the nominative singular, or in the first person singular forms respectively. Compounds are parsed into their structural constituents which, according to the Greek language characteristics, are morphemes (i.e., stems and affixes) or words depending on the case. A first approach to stress is also taken into consideration since compounds display peculiar stress properties that differ from other word-stress properties and are related to syllabification techniques. As is shown below, the system is not able to provide a linguistically-sound analysis for phenomena such as stress and syllabification

Manuscript received in 2004.

This work was co-funded by 75% from E.E. and 25% from the Greek Government under the framework of the Education and Initial Vocational Training Program – Archimedes.

Angela Ralli is with the department of Philology, Division of Linguistics, University of Patras, GR-265 00, Rio, Patras, Greece (e-mail: ralli@upatras.gr; aralli@cc.uoa.gr).

Eleni Galiotou is with the Department of Informatics, Technological Educational Institute of Athens, Ag. Spyridona, GR-122 10 Egaleo, Greece (e-mail: egali@di.uoa.gr; egali@teiath.gr).

II. The Basic Characteristics of Parsing with PC-KIMMO

In the two-level model, as has been formulated by Koskenniemi ([8]) and Karttunen ([6]), a word is taken to be a mapping between a lexical form and its surface (textual) form. The model distinguishes between the word's *morphotactics,* that specify its morpheme constituents in the particular order into which they occur, and the word's *morphophonemics* which account for the various orthographic forms of the morphemes.

The *lexicon* incorporating the morphotactics consists of a list of morphemes. Each lexical entry is specified by information referring to grammatical category, morphosyntactic features, a gloss (additional information), and an alternation index indicating the list of alternative morphemes that may be combined with it. Lexical entries are generally grouped into sublexica, depending on their grammatical category. See in (1) the sublexica that are used here for the purposes of our system:

(1) N (noun), V (verb), ADJ (adjective), DET (determiner), P (preposition), PR (pronoun), ADV (uninflected adverb), CONJ (conjunction), IJ (interjection), PART (particle), CLITIC, ADI (inflected adverb), PRI (inflected pronoun), DAF (derivational suffix), PREFIX, SUFFIX, INFL (inflectional ending).

An example of a lexical entry of the sublexicon of nouns is given in (2):

(2) ánθρωπ- [anθrop] "man"

```
;Sublexicon N
\lf ánθρωπ+
\lx N
\alt Suffix
\fea masc 2 ;
\gl N(ánθρωπ+/man)
```

In (2), the sublexicon entry consists of a record comprising the fields denoted by the following codes:

\lf (lexical item): the morpheme at the lexical level. The entry is transcribed according to the characters of the Greek alphabet since the system deals with the orthographic form of Greek words. The first character bears a stress "á" since, according to Ralli [10], stems and derivational affixes bear

stress properties, as opposed to inflectional affixes that are deprived of such properties.

\lx (sublexicon): the grammatical category

\alt (alternation): a slot containing the list of the continuation classes i.e., the grammatical categories of the morphemes that may follow during word formation.

\fea (features): a list of associated features. In (2), the abbreviation masc stands for the attribute-value pair gender=MASC and the abbreviation 2 for the inflection class (ic) of the entry which is expressed as an attribute-value pair ic=2. As shown by Ralli ([12], [13], [15], [16]), gender and inflection class are features inherent to nominal stems.

In its original conception ([8]), the two-level model segments the word in its constituent parts, and accounts for word-internal phonology and orthography. In this model, the rules remind of the rules of SPE generative phonology, but also differ from them in several aspects. SPE rules are unidirectional and are applied sequentially, converting underlying forms to surface forms through a number of intermediate levels of representation. As opposed to SPE rules, two-level rules are declarative, expressing correspondences that hold between a lexical and a surface form, apply in parallel, and do not allow any intermediate levels of representation. Because of their relational character (i.e., they represent correspondences between surface and lexical forms) they are bi-directional. Two-level rules are implemented as Finite State Transducers and their associated machines run in parallel. A Finite State Transducer (FST) is a kind of Finite State Automaton (FSA) that is composed of a set of nodes, called states, and a set of labeled directed transition arcs. One state is characterized as the initial state and there are one or more final states. An FSA operates on an input string. A successful transition from one state to another is possible when the current symbol of the input string matches the symbol on the arc connecting the states. If the input string is exhausted and the machine is in a final state, the input string is recognized by the FSA. An FST functions like an FSA, but it operates on two input strings and the label on the arc of an FST consists of a valid correspondence pair of symbols of the two input strings.[1]

Koskenniemi ([8]) has proposed a grapho-phonological processing model where all rules apply simultaneously and where each rule can be compiled into an FST. The two-level rules are represented as state transition tables, the rows of which represent the states of the FST where the number of a final state is marked with a colon and the numbers of the non-final states are marked with a period. The columns represent the arcs from one state to another and the column headers are pairs of symbols.In our system, the phonological rules follow the principles of lexical phonology and are generally applied at

morpheme boundaries. For instance, the rule in (3) describes the correspondence between a character 'χ' at the lexical level and a character 'ξ' at the surface level, before a character 'σ' which is realized as a surface '0' (i.e., it is deleted). The correspondence holds at a morpheme boundary (+) that is also realized as a surface 0.

(3) RULE "χ:ξ => __+:0 σ:0" 3 4
χ + σ @
ξ 0 0 @
1: 2 1 1 1
2. 0 3 0 0
3. 0 0 1 0

In (3) above, the symbol '=>' denotes that the correspondence is valid "only but not always" in the environment described by the rule. This rule accounts for the change of the stem-final consonant 'χ' ([x]) into a 'κ' ([k]), before the 'σ' ([s]) that marks the "perfective" aspectual value of verbal types such as *`etreksa* "run-PERF-1P-SG" (4). It should be noticed, however, that in orthographic terms, the consonant cluster [ks] is written as 'ξ', that is why 'κ' ([k]) does not appear on the rule.

(4) *`trexo*[2] < *trex* *o*
I run run IMP-1-SG[3]
vs.
`etreksa < e *trex* *sa*[4]
I ran run-PERF-PAST-1-SG

In addition, the rule file contains a list of stress rules which are discussed in section III.B, as well as an epenthesis rule which is discussed in section III.D.

The implementation in LISP of two-level morphology has been realized by Karttunen ([6]), and named KIMMO after Koskenniemi's first name. PC-KIMMO version 1 is closely related to KIMMO. It has been developed at the Summer Institute of Linguistics (see Antworth 1990) and implemented in C. Originally, the system could tokenize a word into a sequence of tagged morphemes, but could not directly determine its grammatical category and/or its inflectional features. In order to remove this deficiency, and allow PC-KIMMO to act as a morphological front-end to a syntactic parser, a third component has been added that is a unification-based chart parser, following the PATR-II formalism (see [19]). PC-KIMMO version 2 that is used for the purposes of our work handles a word grammar which has the power of a context-free grammar and can model word structures as arbitrarily complex branching trees (see [2]). Thus, when a word is submitted to recognition, it is tokenized into a sequence of morpheme structures by the *rules* and the *lexicon*. The result of the analysis is passed to the *word*

[1] See [5] and [7] for the description of an application of FST's in Computational Linguistics.

[2] For the purposes of this paper, Greek words have been transcribed according to the characters of The International Phonetic Alphabet. For typographical reasons, when necessary, stress is indicated with the symbol " ` " before stressed syllables.

[3] The glosses stand for imperfective (IMP), perfective (PERF), past tense (PAST), first person (1P), singular (SG).

[4] *e-* is the augment, that is a prefix-type element that is a stress carrier and is inserted before verbal stems in the past tense.*-sa* is the inflectional ending carrying the values of perfective *(-s),* past, first person, singular *(-a*).

grammar which returns a parse tree and a feature structure. A feature structure is associated with each node of the parse tree, while the one associated to the top node contains the features that are attributable to the whole word. The implementation of a word grammar dealing with the properties of compound formation is described in sections III.A. and III.C.

III. INCORPORATING PROPERTIES OF GREEK COMPOUNDS IN PC-KIMMO, VERSION 2

Modern Greek is particularly rich in compound formations. According to Ralli ([11], [13]), they are usually defined as an association of two stems or of a stem and a word. Compounds occur as one unit on phonological, morphological, syntactic and semantic grounds and display the following characteristics:

- A Greek compound constitutes one phonological word since it bears only one stress that may be independent of the stress of its constituent units when used as separate words.
- Nominal or verbal compounds are always inflected at their right edge and do not bear word-internal inflection. The absence of internal inflection follows from the fact that nominal or verbal constituents that are used as first members in compound formations belong to the morphological category of stems. That is they are deprived of their inflectional endings and, as such, cannot appear as independent words. In case that a word occurs as the first member of a compound, it is always an uninflected one, as shown below.
- Compounds have an atomic character. That is syntactic principles and operations do not affect their word-internal structure. For instance, compounds are anaphoric islands, and their structural contiguity cannot be interrupted by the insertion of elements functional or non-functional, e.g., adjectives, adverbs, clitics, determiners, prepositions, etc.
- The meaning of compounds is rarely fully compositional. It is driven by the necessity to form new concepts and is produced on the basis of more elementary ones, that is on the basis of the meanings of their constituent parts. As is the case of morphological expressions (cf. [3]), the semantic result of compounding is the creation of a new entity which is interpreted attributively/generically.

A. Constituent structure of Greek Compounds

Greek compounds generally belong to the major grammatical categories of nouns, adjectives and verbs.[5] They are built from constituents each belonging to one of the categories noun, verb and adjective. In what follows, we list representative examples of the most frequent Greek compound types in a broad phonetic transcription. Nouns and adjectives are given by convention in the nominative singular form of the masculine gender type while verbs are cited in the first person singular of the present tense. Inflectional endings appear in parentheses. Absence of parentheses denotes zero inflectional endings.[6]

(5) a. N < N + N

laxanaγo`ra	<	*laxan-*	*aγora*
vegetable market		vegetable	market
xar`tokut(o)	<	*xart-*	*kut(i)*
paper box		paper	box
kozmoxala`zm(os)	<	*kozm-*	*xalazm(os)*
world destruction, chaos		world	destruction

b. N < A + N

aγri`anθrop(os)	<	*aγri-*	*anθrop(os)*
wild man		wild	man
ela`fropetra	<	*elafr-*	*petra*
light stone, pumice		light	stone

c. A < A + A

a`spromavr(os)	<	*aspr-*	*mavr(os)*
white (and) black		white	black
ikonomikopoliti`k(os)	<	*ikonomik-*	*politik(os)*
economic (and) political		economic	political
mikrokamo`men(os)	<	*mikr-*	*kamomen(os)*
small made, small		small	made

d. A < ADV + A

aδikoxa`men(os)	<	*aδik-*	*xamen(os)*
lost in vain		in vain	lost
arγo`kinit(os)	<	*arγ-*	*kinit(os)*
slow moving		slow	moving

e. A < N + A

kozmo`ksakust(os)	<	*kozm-*	*ksakust(os)*
world famous		world	famous
iljoka`men(os)	<	*ilj-*	*kamen(os)*
sunburnt		sun	burnt

f. V < N + V

xarto`pez(o)	<	*xart-*	*pez(o)*
play cards		card	play
xaropa`lev(o)	<	*xar-*	*palev(o)*
fight (with) death		death	fight-V

g. V < V + V

pijeno`erx(ome)	<	*pijen-*	*erx(ome)*
come (and) go		go	come

[5]As claimed in [11], the category of adverbs results from a suffixation process that adds the most common adverbial suffix *–a* to compound formations of a nominal category.

(i)a. *vorioanatoli`ka* < [[[*vori*] o [*anatolik*]] *a*]
northeast-ADV [north east]-ADJ -ADV

b. *meso`ximona* < [[[*mes*] o [*ximon*]] *a*]
mid-winter-ADV [mid winter]-N -ADV

[6] See [14] for a detailed analysis of Greek inflection and inflection classes.

aniγo`klin(o) < *aniγ-* *klin(o)*
open (and) close open close

h. V < ADV + V
kakometaxi`riz(ome) < *kak-* *metaxiriz(ome)*
badly treat badly treat
stravopa`t(o) < *strav-* *pat(o)*
miss one's footing awkwardly step-V
ksana`vrisk(o) < *ksana* *vrisk(o)*
re-find again find

i. N or A < UNINFLECTED ITEM (numeral, pronoun, adverb) + N or A
eγokendri`kos < *eγo* *kendrik(os)*
egocentric self central
e`ksoporta < *ekso* *porta*
out-door out door

As described in Ralli ([11], [13]), most Greek compounds are endocentric and right-headed. The basic patterns generating their word-formation structure are given in (6), and are motivated on phonological and morphological grounds.

(6) a. [Stem Stem], e.g., *xar`tokuto* "paper box" (5a)
b. [Stem Word], *laxanaγo`ra* "vegetable market" (5a)
c. [Word Word], *ksana`vrisko* "re-find, find again" (5h)
d. [Word Stem], *e`ksoporta* "out-door" (5i)

In a context-free grammar that is required by PC-KIMMO, these morphological patterns are generated by a set of context-free rules corresponding to the following fragment of word grammar:

(7) a. Stem -> STEM STEM (generating pattern 6a: [Stem Stem])
Stem -> NWORD STEM (generating pattern 6d: [Word Stem])
Word -> Stem INFL (general word-formation rule generating inflected words containing a non-terminal stem)

b . Word_1 -> NWORD Word_2 (generating pattern 6c: [Word Word])
Word_1 -> STEM Word_2 (generating pattern 6b: [Stem Word])
Word_2 -> STEM INFL (general word formation rule generating inflected words containing a terminal stem)

It should be noticed that in (7) above, Word, Word_1, Word_2 and Stem are non-terminal symbols, while STEM, INFL (inflectional ending) and NWORD (non-inflected word which stands for the uninflected item of (5i)) are the terminal ones.

The context-free rules are enriched with features carrying morphosyntactic information that percolates to the topmost nodes according to the principle of "relativized head" (see [4]). For instance, noun stems are marked for grammatical category and gender (as stated above, gender is a feature inherent to stems, (cf. [14],[16]), inflectional endings are marked for case and number, while both stems and inflectional affixes are characterized by an inflection-class marker (ic) that operates as a matching device between the two, and ensures well-formed inflected words.[7] Grammatical category, gender, case and number percolate to the word node, while percolation of inflection class is prohibited, since this piece of information has a pure morphological value and is not visible to syntax. (8) provides an illustration of the percolation of features handled by a context-free rule generating nominal inflected words. Feature-passing operations are formulated with the use of the unification device, and the rule succeeds only if the ic features of STEM and INFL unify, thus ensuring the well-formedness of inflected words.

(8) Word = STEM INFL
<Word head gcat> = <STEM gcat>
<Word head agr gender> = <STEM gender>
<Word head agr case> = <INFL case>
<Word head agr number> = <INFL number>
<STEM ic> = <INFL ic>

B. Phonological evidence

From a phonological point of view, it has been shown by Nespor and Ralli ([9]) that the position of stress in compounds depends on their constituent structure and the notion of headedness. Compounds belonging to the first and the last type (6a,d) are submitted to a compound-specific law of an antepenultimate-syllable stress. Compounds of the other two types carry the stress of the right-hand head which is a word and a phonological word as well.

For instance, *ku`klospito* and *e`ksoporta* (9a,b) bear the stress on the antepenultimate syllable, that is on a different syllable from the one where the two constituent members are stressed when used separately.

(9) a. *ku`klospito* < *`kukl(a)* *`spit(i)*
doll's house doll house
b. *e`ksoporta* < *`ekso* *`porta*
out-door out door

As Nespor and Ralli claim, compounds like these in (9) contain a stem as head of the construction that does not have any fixed stress properties. That is why they are subject to the

[7]As proposed by Ralli ([14]), Greek nominals are inflected according to 10 inflection classes.

application of a specific compound-stress law according to which, stress falls on the antepenultimate syllable of the formation.

(10) illustrates the implementation of the compound–stress law in our system as a two-level rule. It expresses a correspondence between lexical forms and surface forms containing a stressed antepenultimate syllable, and states that the vowel of the antepenultimate syllable (3rd vowel from the end of the word) is the one carrying the stress. [8]

(10) RULE "V:Vs => __ [C*VC*VC#]" 4 5

	V	V	C	#	@	
	Vs	V	C	#	@	
1:	2	1	1	1	1	
2.	0	3	2	0	1	
3.	0	4	3	0	1	
4.	0	0	0	1	0	

where # represents the word boundary, and C,V and Vs represent the subset of consonants, the subset of unstressed vowels and the subset of stressed vowels of the Greek alphabet respectively.

This correspondence is valid in the particular context, and applies "only but not always". Consequently, it does not block the analysis of words with fixed stress properties.

On the contrary of the (6a,d) cases carrying a stress on the antepenultimate syllable, in the (6b,c) cases, the stress of the compound is the same as the one of the head of the structure, this being a word with fixed stress properties. According to Nespor and Ralli ([9]), a change in the stress is forbidden by a stress-preservation principle that preserves the stress of the head throughout the compound. That is why in (11a) the compound carries the stress of the word *aγo`ra* "market", and in (11b) the stress of the word *`vrisko* "find". It should be noticed that in cases such as (11b), the stress of the first word constituent *ksa`na* "again" is eliminated in favor of the second because Greek compounds constitute phonological words with only one stress. Thus, the second constituent, that is the head, is stronger than the non-head and triggers the stress elimination of the latter.

(11) a. *laxanaγo`ra* < *laxan* *aγo`ra*
vegetable market vegetable market
b. *ksana`vrisko* < *ksa`na* *`vrisko*
re-find again find

Computationally, this situation appears rather complicated due to the limitations of the PC-KIMMO software. According to Antworth ([1]) "Suprasegmental elements such as stress, length and tone must be represented as symbols interspersed with segmental segments of the same level". In an earlier treatment of Greek inflection with PC-KIMMO (cf. [18]), a set of stress operators was adopted which were responsible for determining stress and stress movement in Greek words. This set was defined at the lexical level and was mapped into null (0) symbols at the surface level. Although technically this solution seems to work in a quite satisfactory way, it remains questionable from a linguistically-sound point of view. However, in a first stage of our experimentation, we have also adopted a single stress operator that, in cases like (11) above, blocks the application of the antepenultimate-stress rule, thus applying the stress-preservation principle. We believe that in order to reach a theoretically-elegant approach of stress phenomena in Greek, syllabification has to be accounted for, something which proves to be quite difficult with PC-KIMMO and has not been dealt yet (SIL, personal communication). Therefore, a systematic treatment of stress and syllabification with PC-KIMMO in Greek still remains an open question.

C. Morphological evidence

On morphological grounds, the word-formation structures given in (6) are motivated on the basis of inflection. Inflection appears at the right-hand side of a compound construction, since, as stated above, there is no word-internal inflection. That is why in the structural patterns of (6a,b) a stem always occurs as the first member of the compound. As for the structures in (6c,d) where a word appears to be the first member, it is crucial that it constitutes an uninflected word, that is an adverb, a numeral or an uninflected pronoun.[9] It should be noticed, however, that in spite of the fact that inflection follows the second constituent of the compound structure, it does not directly combine with it in all cases. As opposed to Zwicky's claim ([22]) according to which the head of a morphological construction is defined as the morphosyntactic locus that bears the overt inflectional ending of the construction, in compounds of the first and the fourth type (6a,d), inflection combines with the compound structure as a whole. This is proved by the fact that, most of the times, the inflectional ending following the compound structure is not the same as the one used by the second constituent when used separately:

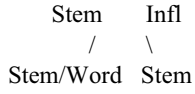
(12) xar`tokut(o) < xart(i) kut(i)
paper-box paper box

Thus, as claimed in [13], compounds like the ones in (12) display the structure of [[Stem Stem] Infl] (6a) or the one of [[Word Stem] Infl] (6d). Both cases are depicted in (13) below under the form of a tree representation.

(13) Word
/ \

[8]In orthography, stress falls on the vowel of the stressed syllable, while here, stress is given before the syllable for technical reasons.

[9]Cases such as *niktilam`pis* < *nikti –lampis* "who shines at night" and *ori`vatis* < *ori –vatis* "who climbs mountains" which contain inflected words like *nikti* "night-DATIVE" and *ori* "mountain-DATIVE" as their first member should be considered to be lexicalized remnants of an Ancient Greek stage of language where word-internal inflection was possible (cf. Ralli 1998 for more details).



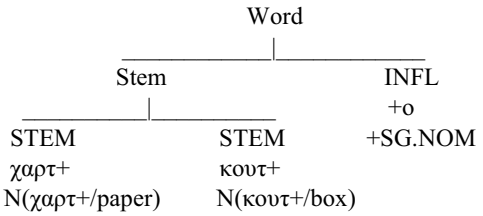
These structures are handled by the rules of the word grammar. As an example, the output of the recognition process for the [[Stem Stem] Infl] structure is depicted in (14) below. Initially, the system delivers the result of the segmentation process, that is a list of morphemes associated to their glosses. Then, the processing is passed on to the word grammar which delivers a parse tree and a feature structure associated to the top node. The feature structure contains information that is the output of the unification process.

These structures are handled by the rules of the word grammar. As an example, the output of the recognition process for the [[Stem Stem] Infl] structure is depicted in (14) below. Initially, the system delivers the result of the segmentation process, that is a list of morphemes associated to their glosses. Then, the processing is passed on to the word grammar which delivers a parse tree and a feature structure associated to the top node. The feature structure contains information that is the output of the unification process.

(14) *χαρτόκουτο* [xar`tokuto] “paper-box”

χαρτ+κουτ++ο N(χαρτ+/paper)N(κουτ+/box)+SG.NOM

1:

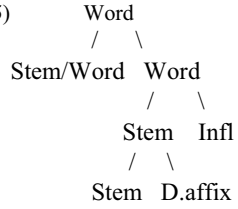


Word:
[mcat: Word
head: [agr: [case: NOM
gender:NEUT
number:SG]
gcat: N]]

1 parse found

It is crucial to note that in the formations of (11) the second member, i.e., the head of the compound structure, is not a derived word. Generally, compounds with a head containing a derivational affix fall under the structures of (6b,c), that is under the [[Stem/Word] [Stem Infl]] structure, since the inflectional ending of the compound is never different from the inflectional ending taken by the derived head, when used as a separate word. The tree representation of (6b,c) is given in (15) below, where the stem of the second member of the compound may be morphologically simple or derived.

(15)

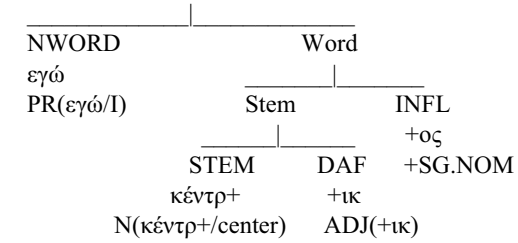


The claim that compounds with derived heads are generated by the structure of (15) is also supported by phonological evidence since they generally preserve the stress of the head, that is of the second member, while the stress of the non-head is eliminated. The antepenultimate-syllable stress law that is typical to compounds of (12) does not apply to formations generated by the tree in (15). In (16) we give the parsing of a compound belonging to the (6c) structure which also contains a derivational affix. In this case, the stress of the uninflected word *εγώ* [e`γo] “I” is eliminated in favor of the stress of its second member *κεντρικός* [kendri`kos] “central”.[10]

(16) *εγωκεντρικός* [eγokendri`kos] "self centered, egocentric"

εγώ κέντρ++ικ+ος
PR(εγώ/I)N(κέντρ+/center)ADJ(+ικ)+SG.NOM

1:



Word:
[mcat: Word
head: [agr: [case: NOM
gender:MASC
number:SG]
gcat: ADJ]]

1 parse found

In the above, the resulting grammatical category of ADJ is imposed by the head, that is by the derivational affix *-ικ-,* represented on the tree as DAF.

An example of a verbal compound which falls under the same (6c) structure, but without any derivational affix is given

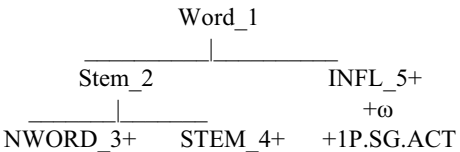
[10]According to Ralli ([10]), in a derived word, the stress is determined by the stress properties of the derivational affix. For instance, the word *kendri`kos* assumes its stress from a lexical phonological property of the derivational suffix *–ik-* which states that stress follows on the first vowel following *–ik-*. It should be noticed that the stress of the stem *kendr-* “centre” is eliminated in favor of the stress properties of *–ik-*, while the inflectional affix *–os* has no particular stress properties. See [20] for a computational analysis of Greek inflected words.

in (17):

(17) *ξαναγράφω* [ksana`γrafo] “re-write”

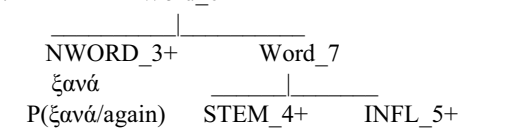
ξανά γράφ++ω
P(ξανά/again)V(γράφω-+/write)+1P.SG.ACT

1:



Word:
[mcat: Word
head: [agr: [number:SG
pers: 1P
tense: PRES
voice: ACT]
gcat: V]]

2:



Word:
[mcat: Word
head: [agr: [number:SG
pers: 1P
tense: PRES
voice: ACT]
gcat: V]]

2 parses found

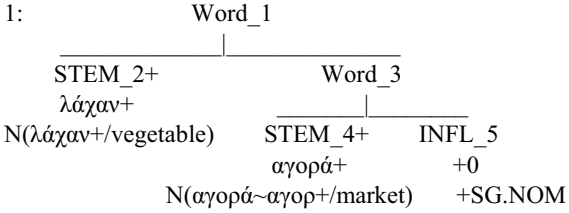
In the word *ξαναγράφω* [ksana`γrafo] “re-write”, the stress preservation principle should block the application of the antepenultimate-syllable stress rule and, thus, eliminate the first parse. However, the absence of a systematic treatment of stress and syllabification, due to the limitations of PC-KIMMO, has led to the over-recognition of the word, and provided the first parse.

In (18), we give the output of a recognition process of compounds falling under (6b):

(18) *λαχαναγορά* [laxanaγo`ra] "vegetable market"

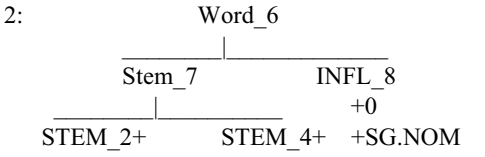
λάχαν+αγορά++
N(λάχαν+/vegetable)N(αγορά~αγορ+/market)+SG.NOM

1:



Word:
[mcat: Word
head: [agr: [case: NOM
gender:FEM
number:SG]
gcat: N]]

2:



Word:
[mcat: Word
head: [agr: [case: NOM
gender:FEM
number:SG]
gcat: N]]

2 parses found

In (18), we see that the system correctly recognizes the case where the head of the construction is a word with fixed stress properties and gives us the first parse. As is the case in (17), the absence of a systematic treatment of stress and syllabification results into the second parse.

D.The linking vowel

As shown in [11], the structure of most compounds contains a linking vowel –o- between the first and the second member. This –o- originates from an ancient thematic vowel that was added to a root. Today, it has lost its stem-bound status since it follows roots that in Ancient Greek were never combined with it. [11]

(19)a. *pa` γovun(o)* < *payice-berg* *γ-* *ice* *vun-* *mountain*
b. *tirosa`lata* < *tir-* *salata*
cheese-salad cheese salad

This vowel is neither a derivational affix, since its only function is to denote a transition between the two members in

[11] See [17] for a diachronic study of the linking vowel.

a compound structure, nor an inflectional affix because it remains unchanged when the morphosyntactic features of case and number denoted by the first member vary according to the context. Examples in (20) provide an illustration to this remark.

(20)a. *a`spr**o**mavr(**o**)*[12] < *aspr(o)* *mavr(o)*
white black
white(and)black-NEUTER-SINGULAR-NOMINATIVE

b. *a`spr**o**mavr(**a**)*
white (and)black-NEUTER-PLURAL-NOMINATIVE
vs.
a`spra**mavr(**a**)*

c. *a`spr**o**mavr(**u**)*
white (and) black-NEUTER-SINGULAR-GENITIVE
vs.
a`spru**mavr(**u**)*

d. *a`spr**o**mavr(**on**)*
white (and) black-NEUTER-PLURAL-NOMINATIVE
vs.
a`spron**mavr(**on**)*

In (20), we see that in a compound denoting a coordinative relation between its two members, the right-hand inflection changes according to the case, while the internal –o- keeps its original form independently of any morphosyntactic features.

Crucially, –o- does not occur in compounds where the first member is a word. That is –o- is absent from compounds of a structure of the (6c,d) type.

(21)a. *ksana`vrisk(o)* < *ksana* *vrisk(o)*
re-find again find
vs.
*ksana**o**`vrisk(o)

b. *e`ptalof(os)* < *epta* *lof(os)*
seven-hill(ed) seven hill
vs.
*epta`**o**lof(os)

Thus, –o- is bound to compound structures where the first member, i.e., the non-head, is a stem[13], and usually appears when the second member, i.e., the head, begins by a consonant. It should be noticed that the presence of a vowel-initial second member triggers the non-occurrence of –o-, as the examples in (22a) show, unless it is the case of a coordinative relation between the members of the compound (see 22b,c).

(22)a. *αγρι`ανθρωπ(ος)* < *αγρι-* *ανθρωπ(ος)*
wild man wild man
vs.
b. **αγρι**o**`ανθρωπ(ος)*

c. *ανγλ**o**αμερι`κ(ος)* < *ανγλ-* *αμερικ(ος)*
anglo-american English American
vs.
d. **ανγλαμερι`κ(ος)*

e. *πηγαιν**o**`ερχ(ομαι)* < *πηγαιν-* *ερχ(ομαι)*
come (and) go go come
vs.
f. **πηγε`νερχ(ομαι)*

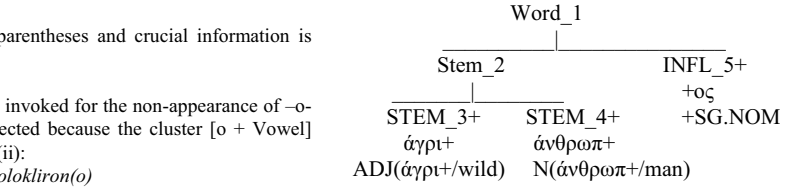
In PC-KIMMO, the linking vowel phenomenon is handled by an epenthesis two-level rule (23) stating that the linking vowel –o- is inserted at the surface level only after a morpheme boundary (+). Thus, it captures the cases in (19). It also states that the linking vowel is not inserted before a vowel, as seen in (22a). Crucially, the rule applies as "only but not always" in the particular context, in a way that it forbids blocking of cases like those in (22c) and (22e).

(23) RULE "0:o => [C | V] +:0 __ C" 4 5
C V + 0 @
C V 0 o @
1: 2 2 1 0 1
2: 2 2 3 0 1
3: 2 2 1 4 1
4. 2 0 0 0 0

In (24), we give the parse tree and the feature structure that are related with the compound *αγριάνθρωπος* "wild man". The two parses found are perfectly legitimate since the head constituent *άνθρωπ(ος)* is already stressed on the antepenultimate syllable and the system cannot block one of the two possible structures.

(24) *αγριάνθρωπος* [αγρι`ανθροπος] "wild man"

άγρι+άνθρωπ++ος ADJ(αγρι+/wild)N(άνθρωπ+/man)+SG.NOM
1:



Word:
[mcat: Word

[12] Inflectional endings are put in parentheses and crucial information is given in bold face characters.

[13] Phonological reasons that may be invoked for the non-appearance of –o- in the examples of (11) should be rejected because the cluster [o + Vowel] occurs elsewhere. See the examples in (ii):
(ii)a. *ksanaolokli`ron(o)* < *ksana* *olokliron(o)*
re-finish again finish
b. *epta`orof(os)* < *epta* *orof(os)*
seven storey(ed) seven storey

```
head:   [ agr:    [ case:  NOM
                    gender:MASC
                    number:SG ]
          gcat:  N ]
```

2:

```
                Word_6
          ___________|___________
      STEM_3+              Word_7
       άγρι+          _______|_____________
  ADJ(άγρι+/wild)   STEM_4+     INFL_5+
                     άνθρωπ+      +ος
            N(άνθρωπ+/man)        +SG.NOM
```

```
Word:
[ mcat:  Word
  head:   [ agr:    [ case:  NOM
                      gender:MASC
                      number:SG ]
            gcat:  N ] ]
```

2 parses found

It should be noticed that if we give an ungrammatical word-form such as **αγριοάνθρωπος,* containing a linking vowel between the two constituents, the system is unable to find any parse.

(25) *αγριοάνθρωπος* [αγριο`ανθρopos] "wild man"
NONE

However, the parsing examples of (26) and (27) below show the inadequacy of the system to handle cases such as (22d,f) which are exceptions to the rule. Normally, the example in (26) should have resulted into a legitimate parse tree as well as an appropriate feature structure, while the example of (27) should have been rejected. As seen below, this is not the case. A possible remedy to the situation would be to mark the particular stems while entering in a coordinative relation, as allowing the presence of the linking vowel. However, the system does not handle features triggering the application of rules. Feature marking can be done only at the lexical-entry level.

(26) *αγγλοαμερικάνος* [angloameri`kanos]
"english-american"
NONE

(27) **αγγλαμερικάνος* [anglameri`kanos]
"english-american"

```
άγγλ+αμερικάν++ος
ADJ(άγγλ+/english)ADJ(αμερικάν+/american)+SG.
NOM
```

1:

```
                        Word
          ________________|________________
       STEM                              Word
        άγγλ+                    _________|_________
  ADJ(άγγλ+/english)          STEM              INFL
                               αμερικάν+         +ος
                        ADJ(αμερικάν+/american)
+SG.NOM
```

```
Word:
[ mcat:  Word
  head:   [ agr:    [ case:  NOM
                      gender:MASC
                      number:SG ]
            gcat:  ADJ ] ]
```

1 parse found

E. Exocentric compounds

According to Williams' right-hand head rule ([21]), the rightmost node in any binary morphological structure will always be the head of the structure. In fact, Greek compounds are subject to this rule, since the grammatical category and the morphosyntactic features of the right-hand constituent, i.e., a stem or a word according to the case (see (6)), are inherited by the compound as a whole. However, Greek has a number of exocentric compounds which do not fall under the general framework of headedness according to which the second member of a compound structure is the head of the construction.

(28)a. *ipsi`lomisθ(os)* < *ipsil- misθ-*
(who earns a) high salary-ADJ high salary

b. *ka`lotix(os)* < *kal- tix-*
(who has) good luck-ADJ good luck

c. *kokino`trix(is)*< *kokin- trix-*
red-ADJ-FEM hair-N-FEM
(who has) red hair-ADJ/N-MASC

Exocentric compounds show a unique behavior with respect to certain points:

- Semantically, the meaning of an exocentric compound does not denote a subset of the entities denoted by the second member of the construction, as opposed to what happens to the meaning of endocentric compounds. For example, *ipsi`lomisθos* (28a) does not designate a salary which is high but rather someone who earns a high salary.
- Exocentric compounds differ from endocentric ones with respect to category and other morpho-syntactic features which are not inherited from any of the constituent parts that may be considered to be the head of the structure. For instance, *kokino`trixis* (28c) may

be used as an adjective of masculine gender while the right-hand member is a feminine noun (*trix-* "hair") and the left-hand member (*kokin-* "red") that modifies the former is an adjective of an underspecified gender.

- The inflectional endings of exocentric compounds are usually different from the inflectional endings of the second member when taken separately:

(29)a. *ka`lotixos* *vs.* *kali* *tixi*
(who has) good luck good luck

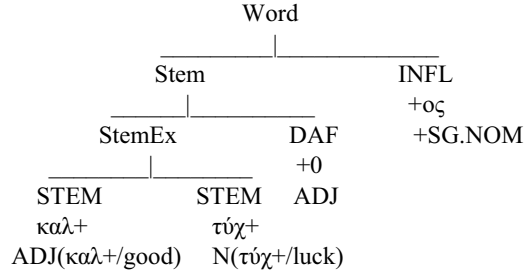
b. *kokino`trixis* *vs.* *kokini* *trixa*
(who has) red hair red hair

Given the last property to display a different inflectional ending from the endings of the constituent parts when used as independent words, Ralli ([11]) has assumed that the basic structure of exocentric compounds follow the (6a) structural pattern, that is exocentric compounds rely on a [Stem Stem] combination schema. Independently of this structural pattern though, they also require an additional device that accounts for their special semantic and morphosyntactic properties. According to [13], they are subject to a conversion process, or to a zero-derivation one, which is responsible for assigning to the compound structure the properties that do not follow from its basic constituent members.

In accordance with the remarks above, in our system, exocentric compounds are handled by the rules of the word grammar which produce outputs such as in (30):

(30) *καλότυχος* [ka`lotixos] "who is lucky"

καλ+τύχ+++ος
ADJ(καλ+/good)N(τύχ+/luck)ADJ+SG.NOM



Word:
[mcat: Word
head: [agr: [case: NOM
gender: MASC
number: SG]
gcat: ADJ]]

1 parse found

In the above, the non-terminal symbol DAF denotes the zero derivational affix and the non-terminal symbol StemEx denotes the formation of an intermediate exocentric stem node. As in the previous examples, the system initially delivers a segmentation of the word into its morphemes at the lexical level. Then, the word grammar delivers the resulting parse and the corresponding feature structure. The ADJ (adjectival) grammatical category is triggered by the zero derivational affix (the DAF node which contains a zero lexical form).

IV. CONCLUSION

In this paper, we have presented the current state of a computational analysis of Greek compounds. We have dealt with the recognition process of nominal and verbal compounds and tried to show certain limitations of the PC-KIMMO software with respect to the principles of compound formation in Greek. We have shown that stress and syllabification that are indispensable for the analysis of such constructions are not handled in a satisfactory way. We believe that a device using feature-triggered rules could remedy this deficiency but, as no modifications to the software are possible at the present stage, we continue our investigation in order to reach more linguistically-acceptable solutions within the particular system. As a future development, we plan to deal with the process of generation, but only when the problems of recognition will be – at least partially - solved.

ACKNOWLEDGMENT

Parts of this paper have been presented at the International Conference of Greek Linguistics (University of Thessaloniki: April 2001) and at the Asymmetry Conference (UQAM: May 2001). We are grateful to the audiences of both conferences for their most constructive remarks.

REFERENCES

[1] E. L. Antworth, *PC-KIMMO: A Two-level Processor for Morphological Analysis, [Occasional Publications in Academic Computing 16],* Dallas, TX : Summer Institute of Linguistics, 1990.
[2] E. L. Antworth, "Morphological parsing with a unification-based word grammar"., in *Proceedings of North-Texas Natural Language Processing Workshop*, University of Texas, Arlington 1994.
[3] A.-M. Di Sciullo, A.-M. 1993. "The complement domain of the head at morphological form". *Probus* 5, 1993 pp. 95-125.
[4] A.-M. Di Sciullo and E. Williams, *On the Definition of the Word.* Cambridge, Mass, MIT Press, 1987.
[5] R. Kaplan and M. Kay, "Phonological rules and finite state transducers". in *Proceedings of the ACL/LSA Conference*, NewYork, 1981.
[6] L. Karttunen, "KIMMO: A General Morphological Processor". *Texas Linguistics Forum 22*, 1983, pp. 163-186.
[7] M. Kay, "When meta-rules are not meta-rules", in *K. Sparck-Jones and Y. Wilks (eds.), Automatic Natural Language Parsing,* Chichester: Ellis Horwood, 1983, pp. 94-116.
[8] K. Koskenniemi, *Two-level Morphology: A General Computational Model for Word-Form recognition and Production* [Publication No. 11, Department of General Linguistics]. University of Helsinki, 1983.
[9] M. Nespor and A. Ralli "Morphology-Phonology interface: Phonological domains in Greek compounds". *The Linguistic Review 13*: 1996, pp. 357-382.
[10] A. Ralli, *Elements de la Morphologie Grecque*. Ph.D. Diss., .University of Montreal, 1998.

[11] A. Ralli, "Compounds in Modern Greek ". *Rivista di Linguistica 4*, 1, 1992, pp. 143-174.
[12] A. Ralli,. "On the morphological status of inflectional features: evidence from Modern Greek". in *G. Horrocks, B. Joseph and I. Philippaki-Warburton (eds.): Themes in Greek Linguistics II*, John Benjamins, 1998, pp. 51-74.
[13] A. Ralli, "Το Φαινόμενο της Σύνθεσης στη Νέα Ελληνική: Περιγραφή και Ανάλυση". ("A description and an analysis of compounding in Modern Greek"). *Parusia* (IA-IB): 183-205. School of Philosophy of the University of Athens, Athens, 1999.
[14] A. Ralli, "A feature-based analysis of Greek nominal inflection". *Glossolojia* (11-12), 2000, pp. 201-227.
[15] A. Ralli, "Gender in Greek Nouns", in *Proceedings of the 4th International Meeting of Greek Linguistics.* Aristotle University Press, Thessaliniki, 2001.
[16] A. Ralli, .*Μορφολογία (Morphology),* Patakis, Athens , forthcoming.
[17] A. Ralli and M. Raftopoulou, "Η Σύνθεση ως Διαχρονικό Φαινόμενο Σχηματισμού Λέξεων" ("Compounding as a diachronic word-formation process"). *Studies in the Greek Language ,* Kyriakides, Thessaloniki, 1999, pp. 389-403.
[18] K. Sgarbas, N. Fakotakis, G. Kokkinakis, "A PC-KIMMO-Based Morphological Description of Modern Greek". *Literary and Linguistic Computing 10*, 3, 1995, pp. 189-201.
[19] S. M. Shieber, *An Introduction to Unification-Based Approaches to Grammar [CSLI Lecture Notes No 4]*. Stanford, CA, 1986.
[20] L. Touratzidis and A. Ralli, "Stress in Greek Inflected Forms: A Computational Treatment". *Language and Speech,* 35, 1992, pp. 435-453.
[21] E. Williams, "On The Notions 'Lexically Related' and 'Head of a Word'". *Linguistic Inquiry 12*, 2, 1981, pp. 245-274.
[22] A. Zwicky, "Heads". *Journal of Linguistics* 21, 1985, pp. 1-29.