2nd Hellenic Conf. on AI, SETN-2002, 11-12 April 2002, Thessaloniki, Greece,
Proceedings, Companion Volume, pp. 53-64

PARSING DEFICIENCIES OF THE PC-KIMMO SYSTEM

Eleni Galiotou[1,2] **Angela Ralli**[3]

[1]Department of Informatics, TEI of Athens, Ag. Spyridona, GR-122 10 Egaleo, Greece
[2]Department of Informatics and Telecommunications, University of Athens, Panepistimiopolis, GR-157 84 Athens Greece
e-mail: egali@di.uoa.gr
[3]Department of Philology, Division of Linguistics, University of Patras, GR-265 00, Rio, Patras, Greece,
e-mail:ralli@upatras.gr

Abstract. In this paper, we discuss the possibilities and limitations of the PC-KIMMO system as a recognition device of compound formations in a language like Modern Greek, where compounding interacts with derivation, inflection and lexical phonology. We deal with the computational processing of nominal and verbal compounds and try to show certain limitations of the PC-KIMMO software with respect to the principles of compound formation. Compounds are parsed into their structural constituents that are morphemes (i.e. stems and affixes) or words, depending on the case. Stress is also taken into consideration since compounds display peculiar stress properties which are different from other word-stress properties. In particular, we show that stress and syllabification that are crucial for the analysis of such constructions cannot be dealt with in a satisfactory way.

1 Morpho-phonological parsing with PC-KIMMO v.2

PC-KIMMO is a morphological parser based on the model of two-level morphology ([10], [11]). 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 different orthographic forms of the morphemes. In its original conception ([10], [11]), the two-level model segments the word in its constituent parts, and accounts for word-internal phonology and orthography by means of declarative *two-level rules* expressing correspondences that hold between a lexical and a surface form. These two-level rules 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. A finite state transducer (FST) functions like a finite state automaton but it operates on

two input strings. The label on the arc of an FST consists of a valid correspondence pair of symbols of the two input strings[1].
The *lexicon* incorporating the morphotactics consists of a list of morphemes. Each lexical entry is characterized by its grammatical category, its morpho-syntactic features, a gloss (additional information), and an alternation index specifying the list of alternative morphemes that may be combined with it. Lexical entries are generally grouped into sublexica, depending on their grammatical category. (1) lists the sublexica used for Greek:

(1) N (noun), V (verb), ADJ (adjective), DET (determiner), P (preposition), PR (pronoun), ADV (non inflected 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) άνθρωπ- [anθrop] «man»

```
;Sublexicon N

\lf άνθρωπ+

\lx N

\alt Suffix

\fea masc 2

\gl N(άνθρωπ+/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
\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. Gender is a feature inherent to nominal stems and ic is also a feature characterizing stems (see [17] for more details on this).

In the grapho-phonological model proposed in [10] and [11] all rules apply simultaneously and each rule can be compiled into an FST. These two-level rules are

[1] See [7] and [9] for the description of an application of FSTs in Computational Linguistics.

represented as state transition tables. The rows of such a table 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.

(3) RULE "φ:ψ => __+:0 σ:0" 3 4

The rule in (3) describes the correspondence between a character 'φ' at the lexical level and a character 'ψ' 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 (+) which is also realized as a surface 0. The symbol '=>' denotes that the correspondence is valid «only but not always» in the environment described by the rule («__+:0 σ:0»). The rile also states that the corresponding state trasition table has 3 rows (states of the FST) and 4 columns (arcs from one state to another). This rule accounts for the change of the stem-final consonant 'φ' [f] into a 'π' [p], before a 'σ' [s] marking the aspectual value of the perfective in verbal types such as *έγραψα* (`eγrapsa) «write-PERF-1P-SG» (4). Notice, however, that the cluster [ps] becomes «ψ» orthographically, that is why 'π'[p] does not appear on the rule.

(4) `γrafo[2] < γraf o
I write write IMP-1P-SG
vs.
`eγrapsa < e γraf sa
I wrote write-PERF-PAST-1P-SG[3]

PC-KIMMO v.1 was developed at the Summer Institute of Linguistics (SIL) and implemented in C ([2]). Originally, the system could tokenize a word into a sequence of tagged morphemes but it 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 unification-based chart parser following the PATR-II formalism ([21]) was added. PC-KIMMO v. 2 ([3]), 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. 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 this analysis is passed to the *word grammar* which returns a parse tree and a feature structure. A feature structure is associated with each node of the parse tree, while the feature structure associated to the top node contains the features that are attributable to the whole word.

[2]Greek words are transcribed according to the characters of the International Phonetic Alphabet. For typographical reasons when necessary, stress is indicated with the symbol « ` « before any stressed syllable.

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

2 Adapting the morpho-phonological parser to the principles of compound formation

Modern Greek is particularly rich in compound formations. According to ([14], [16]), they are usually defined as an association of two stems or of a stem and a word that occur as one unit on phonological, morphological, syntactic and semantic grounds. Consider 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. In case that a word occurs as the first member of a compound, it is always an uninflected one.
- Compounds have an atomic character. That is, syntactic principles and operations do not affect their word-internal structure.
- 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.

Greek compounds generally belong to the major grammatical categories of nouns, adjectives and verbs. They are built from constituents each belonging to one of the categories noun, verb and adjective.

As described in ([14], [16]), most Greek compounds are endocentric (headed by one of their members) and right-headed. The basic morphological patterns generating their structure are the following[4]:

(5) a. [Stem Stem]

e.g.,	xar`tokut(o)	<	xart-	kut(i)
	paper box		paper	box

b. [Stem Word],

e.g.,	laxanaγo`ra	<	laxan-	aγo`ra
	vegetable market		vegetable	market

c. [Word Word],

e.g.,	ksana`γraf(o)	<	ksa`na	`γraf(o)
	rewrite		again	write

d. [Word Stem],

e.g.,	e`ksoporta	<	`ekso	`porta
	out-door		out	door

[4] In the examples of (5) inflectional endings are put in parentheses. Absence of parentheses denotes zero inflectional endings.

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:

(6)a. `Stem -> STEM STEM` (pattern 5a: [Stem Stem])

`Stem -> NWORD STEM` (pattern 5d: [Word Stem])

`Word -> Stem INFL` (general word-formation rule: inflected words containing a non-terminal stem)

b. `Word_1 -> NWORD Word_2` (pattern 5c: [Word Word])

`Word_1 -> STEM Word_2` (pattern 5b: [Stem Word])

`Word_2 -> STEM INFL` (general word formation rule : inflected words containing a terminal tem)

Notice that in (6) above, Word, Word_1, Word_2 and Stem are non-terminal symbols of the grammar while STEM, INFL (inflectional ending) and NWORD (non-inflected word as in (5d)) are the terminal ones.
These context-free rules are enriched with featurized information. In the case of nouns, for instance, stems are marked for gender (as stated above, gender is a feature inherent to stems, cf. [17]) and inflectional endings are marked for case and number. 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.[5] (7) provides an illustration of the percolation of features handled by a context-free rule generating nominal inflected words:

(7) `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>`

[5]As proposed in [17], Greek nominals are inflected according to 10 inflection classes.

It should be noticed that this rule succeeds only if the ic features of STEM and INFL unify. Feature-passing operations such as percolation of category, gender, case and number are all formulated with the use of the unification device. The reason for postulating the patterns in (5a-d) is of phonological and morphological nature.

From a phonological point of view, it has been shown in [13] that the stress of compounds depends on their constituent structure and on the notion of headedness. Compounds belonging to the first and the last type (5a,d) are submitted to a compound-specific law of an antepenultimate-syllable stress. Compounds of the other two types (5b,c) carry the stress of the right-hand head which is a word and a phonological word as well.

For instance, *e`ksoporta (5d)* 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.

As claimed in [13], compounds like these in (5d) contain a stem as head of the construction that does not have any fixed stress properties. That is why they are subject to the application of a specific compound-stress rule according to which, stress falls on the antepenultimate syllable of the formation.

In our system, this compound-stress rule is implemented as a two-level rule which expresses the correspondence between lexical forms and surface forms containing a stressed antepenultimate syllable. It simply states that the vowel of the antepenultimate syllable (3rd vowel from the end of the word) is stressed. In (8) below, # 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. Notice that, this correspondence is also valid «only but not always» in this context so as not to block the analysis of words with fixed stress properties.

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

On the contrary of the (5a,d) cases with a stress on the antepenultimate syllable, in the (5b,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 [13], 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 (9a) the compound carries the stress of the word *αγο`ra* «market», and in (9b) the stress of the word *`γrafo* «write». Notice that in cases such as (9b), 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.

(9) a. laxanaγo`ra < laxan aγo`ra
vegetable market vegetable market

b. ksana`γrafo < ksa`na `γrafo
rewrite again write

Computationally, this situation is rather complicated due to the limitations of the PC-KIMMO software. According to ([2]: page 12) «Suprasegmental elements such as stress, length and tone must be represented as symbols interspersed with segmental segments at the same level». In an earlier treatment of Greek inflection with PC-KIMMO, ([20]), 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. In the first stage of our experimentation a single stress operator that, in cases like (9) above, blocks the application of the antepenultimate-stress rule, would suffice. However, 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). A systematic treatment of stress and syllabification with PC-KIMMO in Greek remains an open question. On morphological grounds, the structural patterns given in (5) are motivated on the basis of inflection. Inflection appears at the right-hand side of a compound construction, because, as stated above, there is no word-internal inflection in Greek compounds. Consider the example in (10) below which falls under the pattern 5c:

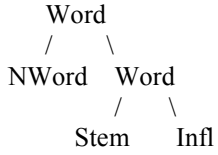
(10)

ksana`γraf(o)	<	ksa`na	`γraf(o)
rewrite		again	write

[NWord [Stem Infl]] (NWord : Non-inflected word)

A tree representation of the compound is as in (11):

(11)



In our implementation, the output of the recognition process of compounds like the one in (10) is as in (12):

(17) *ξαναγράφω* [ksana`γrafo] «re-write»

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

1:                      Word_1
                  ___________|___________
                 Stem_2                  INFL_5+
            _____|_____                   +ω
  NWORD_3+        STEM_4+                 +1P.SG.ACT
     ξανά         γράφ+
 P(ξανά/again)    V(γράφ+/write)
```

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


2:                 Word_6
              ___________|__________
          NWORD_3+              Word_7
            ξανά              _______|________
       P(ξανά/again)        STEM_4+        INFL_5+
                             γράφ+           +ω
                         V(γράφ+/write)   +1P.SG.ACT

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

2 parses found
```

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 which is associated to the top node. 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.

Yet, the stress preservation principle can only be applied at the level of the word structure which is not possible within PC-KIMMO. Therefore, the absence of a systematic treatment of stress and syllabification, due to the limitations of the software, has led to the over-recognition of the word, and provided two parses.

3 The linking vowel phenomenon

As shown in [14], the structure of most compounds contains a linking vowel -o- between the first and the second member. This vowel is neither a derivational affix, since its only function is to denote a transition between the two members in 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. 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.

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

(13)a. aγri`anθrop(os) < aγri- anθrop(os)
wild man wild man
vs.
*aγri**o**`anθrop(os)

b. angl**o**ameri`kan(os) < angl- amerikan(os)
anglo-american English American
vs.
*anglameri`kan(os)

In PC-KIMMO, the linking vowel phenomenon is handled by an epenthesis two-level rule:

(14) RULE "0:o => [C | V] +:0 __ C" 4 5

The epenthesis rule states that the linking vowel –o- is inserted at the surface level only after a morpheme boundary (+). The linking vowel is not inserted before a vowel, as seen in (13a). Notice that the rule is also applied «only but not always» in the particular context so as not to block cases like those in (13b).

(15) αγριάνθρωπος [aγri`anθropος] «wild man»

```
αγρι+άνθρωπ++ος      ADJ(αγρι+/wild)N(άνθρωπ+/man)+SG.NOM

1:                              Word_1
                     ___________|_________________
                   Stem_2                      INFL_5+
                ________|______                  +ος
             STEM_3+          STEM_4+         +SG.NOM
               αγρι+            άνθρωπ+
          ADJ(αγρι+/wild)   N(άνθρωπ+/man)

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

```
2:
                Word_6
          ___________|___________
     STEM_3+                  Word_7
      αγρι+                   ______|_____________
 ADJ(αγρι+/wild)              STEM_4+      INFL_5+
                              ανθρωπ+        +ος
                    (άνθρωπ+/man)            +SG.NOM

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

2 parses found
```

Notice that in the particular case of the compound *αγριάνθρωπος* «wild man», the two parses are 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. The two parses will still be legitimate even after a possible implementation of syllabification in the future.

If we give an ungrammatical word-form such as **αγριοάνθρωπος* with a linking vowel between the two constituents, the system does not find any parsing.

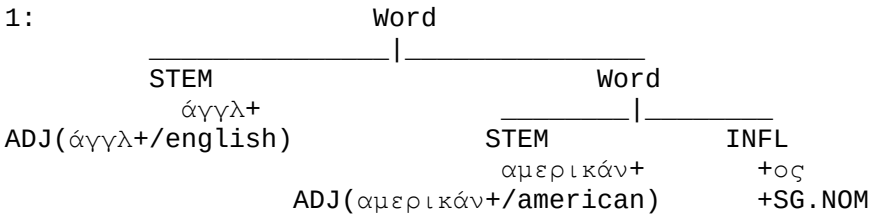
(16) αγριοάνθρωπος [agrio`anθropos] «wild man»
NONE

However, the parsing examples of (17) and (18) below represent the inadequacy of the system to handle cases such as (13b) which constitute an exception to the rule. Normally, the example in (17) should have resulted into a legitimate parsing tree as well as an appropriate feature structure, while the example of (18) should be rejected. As seen below, this is not the case. A possible remedy to the situation would be to mark the stems participating in a coordinative relation but the system does not allow features which trigger the application of rules. Feature marking can be done only at the lexical-entry level, something which is rather awkward from a linguistic point of view.

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

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

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




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

1 parse found
```

6 Conclusions

In this paper, we have discussed on the possibilities and limitations of the PC-KIMMO software with respect to principles of compound formation in Modern Greek. We have dealt with the recognition process of nominal and verbal compounds and we have shown that phenomena such as stress and syllabification which are indispensable for the analysis of such constructions could not be handled in a satisfactory way. We believe that a device using feature-triggered rules could remedy certain deficiencies but, as no modifications to the software are possible, we continue our investigation in order to reach more linguistically acceptable solutions within the particular system

References

1. Anderson, S. R.: Morphology as a Parsing Problem. UCLA Occasional Papers #7: Working Papers in Morphology. University of California, Department of Linguistics, Los Angeles (1988)
2. Antworth, E.: PC-KIMMO: A Two-level Processor for Morphological Analysis, Occasional Publications in Academic Computing No. 16. Summer Institute of Linguistics, Dallas TX (1990)
3. Antworth, E.: Morphological Parsing with a Unification-Based Word Grammar. Proceedings of the North-Texas Natural Language Processing Workshop. University of Texas, Arlington (1994)
4. Di Sciullo, A.M : The Complement Domain of the Head at Morphological Form. Probus, 5. (1993) 95-125.
5. Di Sciullo, A.M. , Ralli A : Argument Structure and Inflection in Compounds: Some Differences between English, Italian and Modern Greek. Proceedings of the Third Meeting

on the Multilingual Aspects of Nominal Composition. ISSCO, University of Geneva, Geneva (1994) 61-76
6. Di Sciullo, A.M. and A. Ralli : Theta-role Saturation in Greek Compounds. In: A. Alexiadou, A., Horrocks, G., Stavrou, M. (eds.) Studies in Greek Syntax. Kluwer, Amsterdam (1999) 185-200
7. Kaplan, R.M., Kay M.: Phonological Rules and Finite State Transducers. ACL/LSA Conference. NewYork (1981)
8. Karttunen, L.: KIMMO: A General Morphological Processor. Texas Linguistics Forum **22** (1983) 163-186.
9. Kay, M.: When Meta-Rules are not Meta-Rules, In: Sparck-Jones K., Wilks Y. (eds.): Automatic Natural Language Parsing. Ellis Horwood, Chichester (1983) 94-116
10. Koskenniemi, K.,: Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. Publication No. 11, University of Helsinki, Department of General Linguistics (1983)
11. Koskenniemi, K.: Two-level Model for Morphological Analysis. Proceedings of IJCAI'83. (1983) 683-85.
12. Markopoulos, G.: A Two-level Description of the Greek Noun Morphology with a Unification-based Word Grammar. In: Ralli, A., Grigoriadou M., Philokyprou, G., Christodoulakis D., Galiotou E. (eds.): Working Papers in Natural Language Processing. Diavlos, Athens 1997
13. Nespor, M. and A. Ralli: Morphology-Phonology Interface: Phonological Domains in Greek Compounds. The Linguistic Review **13** (1996) 357-382.
14. Ralli, A.: Compounds in Modern Greek. Rivista di Linguistica **4** , 1 (1992) 143-174
15. Ralli, A. : 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) 51-74
16. Ralli, A. : Το φαινόμενο της σύνθεσης στη Νέα Ελληνική: Περιγραφή και ανάλυση (A description and an analysis of compounding in Modern Greek). Parousia, IA-IB. University of Athens, School of Philosophy (1999) 183-205
17. Ralli, A.: A Feature-Based Analysis of Greek Nominal Inflection. Glossolojia (2000).
18. Ralli A. and Galiotou E.: Υπολογιστική επεξεργασία των συνθέτων της Νέας Ελληνικής (Computational Processing of Compounds of Modern Greek). Studies in Greek Linguistics 2001. Kyriakidi, Thessaloniki (2001)
19. Ralli A. and Raftopoulou M.: «Η σύνθεση ως διαχρονικό φαινόμενο σχηματισμού λέξεων (Compounding as a diachronic phenomenon of word formation)», Studies in Greek Linguistics 1999. Kyriakidi, Thessaloniki (1999) 389-403
20. Sgarbas, K., Fakotakis, N. Kokkinakis, G.: A PC-KIMMO-Based Morphological Description of Modern Greek. Literary and Linguistic Computing, **10**, 3 (1995) 189-201
21. Shieber, S. M.: An Introduction to Unification-Based Approaches to Grammar. CSLI Lecture Notes, No 4. Stanford, CA. (1986).
22. Williams, E.: On the notions 'lexically related' and 'head of a word'. Linguistic Inquiry **12**, 2 (1981) 245-274
23. Zwicky, A.: Heads. Journal of Linguistics **21** (1985) 1-29