World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:6, 2007

Morpho-Phonological Modelling in Natural Language Processing

Eleni Galiotou, and Angela Ralli

***Abstract*—**In this paper we propose a computational model for the representation and processing of morpho-phonological phenomena in a natural language, like Modern Greek.

We aim at a unified treatment of inflection, compounding, and word-internal phonological changes, in a model that is used for both analysis and generation.

After discussing certain difficulties cuase by well-known finite-state approaches, such as Koskenniemi's two-level model [7] when applied to a computational treatment of compounding, we argue that a morphology-based model provides a more adequate account of word-internal phenomena. Contrary to the finite state approaches that cannot handle hierarchical word constituency in a satisfactory way, we propose a unification-based word grammar, as the nucleus of our strategy, which takes into consideration word representations that are based on affixation and [stem stem] or [stem word] compounds. In our formalism, feature-passing operations are formulated with the use of the unification device, and phonological rules modeling the correspondence between lexical and surface forms apply at morpheme boundaries.

In the paper, examples from Modern Greek illustrate our approach. Morpheme structures, stress, and morphologically conditioned phoneme changes are analyzed and generated in a principled way.

***Keywords*—** Morpho-Phonology, Natural Language Processing.

I. Introduction

THE representation and processing of morpho-phonological knowledge plays a major role in natural language processing, speech processing and information retrieval applications. In this respect, several formalisms have been proposed over the past decades. Undoubtedly, the most successful among them is Koskenniemi´s two-level model [7, 8], which was introduced in 1983 and has been used for the computational processing of morpho-phonological phenomena in a number of languages [5, 16]. One of the major advantages of this model is its use of finite state transducers for the implementation of phonological rules. In fact, the use of finite state methods in computational morphology and other fields of natural language processing has increased, and many substantial results have emerged on theoretical and practical grounds [6]. Although the two-level model has been the most widely accepted practice for morphological processing in the last twenty years, certain difficulties render it inadequate for an efficient unified treatment of both morphological and phonological phenomena in natural language.

In this paper, we take as a starting point certain difficulties that emerged in the computational treatment of Modern Greek compounds. Then, we present our approach, which accounts for a unified treatment of inflection, compounding, and word-internal phonological changes.|

II. A Finite State Approach to the Treatment of Greek Compound Words

A. Basic characteristics of the two-level model

In its original conception [7,8] the model of two-level morphology segments a word in its constituent parts and accounts for word-internal phonology and orthography. Phonological "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, they are bi-directional. The processing is grapho-phonological, all rules apply simultaneously, and each rule can be compiled into a finite state transducer (FST). It incorporates a purely finite-state model of morphotactics, which cannot handle phenomena like long distance dependencies [18]. The first implementation of the model in LISP [5] was named KIMMO while PC-KIMMO v.1, a closely related implementation in C, was developed at the Summer Institute of Linguistics [1]. 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, word syntax had to be taken into account. PC-KIMMO v.2 [2] incorporates a unification-based chart parser, which follows the PATR-II formalism [17].

The word grammar has the power of a context-free grammar, and can model word structures as arbitrarily complex branching trees [2]. Thus, when a word is subject to recognition, it is tokenized into a sequence of morpheme structures by the phonological rules and the lexicon. The result of the analysis is passed to the word grammar, which returns a parse tree and a feature structure. A feature structure

Manuscript received in 2004.

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).

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).

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. In the following we briefly describe our attempt to analyze compounds in Modern Greek (hereafter Greek) with the use of PC-KIMMO v.2, and we discuss the problems that we had encountered [3, 15].

B. *Applying the two-level model to the analysis of Greek compounds*

Greek compounding is of a particular interest, theoretically and computationally, since it interacts with inflection and phonology to a significant extent. To our knowledge, our work, described in [3] and [15], is the first computational work dealing with their internal structure. According to Ralli [10, 11], compounds act as one unit on phonological, morphological, syntactic and semantic grounds. They 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 on their right edge and do not bear word-internal inflection.
- Syntactic principles and operations do not affect the internal structure of a compound.
- The meaning of compounds is rarely fully compositional.

Greek compounds generally belong to one of the major grammatical categories of nouns, adjectives and verbs. According to Ralli [10, 12], the basic patterns generating their structure are motivated on phonological and morphological grounds, and are shown in Figure 1.

a. *[Stem Stem]*
e.g. *xar`tokuto*[1] < *xar`t(i)* *ku`t(i)*
"paperbox" paper box

b. *[Stem Word]*
e.g. *laxanaγo`ra* < *`laxan(o)* *aγo`ra*
"vegetable market" vegetable market

c. *[Word Word]*
e.g. ksana`vrisko < ksa`na `vrisk(o)
"re-find" again find

d. *[Word stem]*
e.g. *e`ksoporta* < *`ekso* *`porta*
"out-door" out door

Figure 1: Basic patterns of compound word–formation

[1]For the purpose of the paper, Greek words are broadly transcribed according to the characters of the International Phonetic Alphabet. For typographical reasons, when necessary, stress is indicated with the symbol " ` " before stressed syllables

The context-free rules generating these patterns correspond to the following fragment of a word grammar, as required by PC-KIMMO:

Stem → STEM STEM
Stem → NWORD STEM
(*NWORD: non-inflected word*)
Stem → STEM DAF
(*DAF: derivational affix used in the derivation process*)
Word → Stem INFL
(*general word-formation rule, generating inflected words containing a non-terminal stem*)
Word_1 → STEM Word_2
Word_2 → STEM INFL
(general word-formation rule, generating inflected words containing a terminal stem)

Figure 2: Fragment of the word gramma in PC-KIMMO format

These context-free rules are enriched with features that carry morpho-syntactic information, which percolates to the word node. Figure 3 provides an illustration of feature-passing operations in nominal inflected words where grammatical category, gender, case, and number are inherited by the word node.

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>

Figure 3: Example of feature passing operations

It should be noticed that according to Ralli [12, 14] stems and inflectional affixes are characterized by an inflectional class marker (ic) that operates as a matching device between the two, and ensures the well-formedness of inflected words. In our approach, phonological rules follow the principles of lexical phonology and are generally applied at morpheme boundaries. For instance, the following rule describes the correspondence between a character 'χ' at the lexical level, and a character 'ξ' at the surface level, before a character 'σ' at the lexical level, which, in turn, is realized as a surface 0 (null). The correpondence holds at a morpheme boundary (+),

which is also realized as a surface 0. This rule accounts for the change of a 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 (I ran) "run-PERF-1P-SG". Note, however, that in orthographic terms, the consonant cluster [ks] is written as 'ξ', that is why 'κ' ([k]) does not appear on the rule.

RULE "χ:ξ => __+:0 σ:0" 3 4

The rule file contains a list of stress rules, which bring up the inadequacy of the two-level approach to treat stress phenomena and will be discussed in the next subsection.

C. Difficulties with the finite state approach

In dealing computationally with Greek compounds, we came across with certain problems that were mostly due to the use of the finite state approach. These problems are extensively discussed in [3] and [15], but are briefly presented here:

- Stress

According to Antworth [1], "Suprasegmental elements such as stress, length, and tone must be represented as symbols interspersed with segmental elements of the same level". Thus, stress and stress movement are represented with the use of stress operators at the lexical level, which are mapped into null (0) symbols at the surface level. This solution proves to be quite problematic in Greek since, in order to have a principled treatment of the antepenultimate stressing procedure, syllabification has to be accounted for, something that is not possible with the PC-KIMMO software [3, 15].

- Feature marking

In the present situation, the word grammar provided by PC-KIMMO offers a more powerful model of morphotactics, which constitutes a deviation from a purely finite state approach, since it has the power of a context-free grammar, and allows a representation of the word structure as an arbitrarily complex branching tree. Moreover, feature structures and the unification mechanism allow for feature passing operations and ensure the well-formedness of words. However, as phonology and word grammar are two modules that function independently from one another, no direct link can be established between them. Therefore, it is not possible to define phonological rules, which are conditioned by the structural properties of word constituents.

In what follows, we propose a modification of the two-level model, which matches the linguistic insights.

III. TOWARDS A UNIFIED TREATMENT OF INFLECTION AND COMPOUNDING

The nucleus of our approach is a unification-based word grammar, similar to the grammar used in the two-level approach. Indeed, a finite state. morphotactics is suitable only to the treatment of inflection. In order to deal successfully with compounding, we need the power of a context-free grammar. Moreover, a unification device is used for feature-passing operations, which is responsible for the determination of the grammatical category of the word and the attribute of the appropriate features. For instance, noun stems are marked for grammatical category and gender (according to Ralli [13], gender is a feature inherent to stems), inflectional endings are marked for case and number, while both stems and inflectional endings are characterized by an inflectional class marker.

Our deviation from the two-level approach focuses on the relationship between the word grammar and the phonological component. As we have already pointed out in section 2, phonological rules apply at morpheme boundaries, and are activated during the word-formation process. As for stressing, it is accounted for after the completion of the word formation process.

A. Syllabification and stress

As shown by Nespor and Ralli [9], a significant number of compounds are built on a [Stem Stem] basis, and are submitted to a compound-specific law of an antepenultimate-syllable stress, while other compounds carry stress on their right-hand component. Take for example the words

ku`klospit(o)* < *`kukl(a) *`spit(i)*
"doll´s house" doll house

e`ksoporta* < *`ekso *`porta*
"out-door" out door

According to [9], these compounds contain a stem as head of the construction, which does not have any fixed stress properties. That is why they are subject to a specific compound-stress rule, according to which, stress falls on the third syllable form the end of the formation. Obviously, a systematic treatment of stress and syllabification is needed separately from segmental phonological rules.

B. Feature marking and feature-triggered rules

As shown by Ralli [10], 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, and is neither a derivational nor an inflectional affix. The linking vowel is bound to compound structures where the first member is a stem and usually appears when the second member begins by a consonant.
Take for instance,

*tir**o**sa`lata* < *tir-* *sa`lata*
"cheese-salad" < cheese salad

Note that the presence of a vowel-initial second member triggers the non-occurrence of the linking vowel, as in the following example,

aγri`anθropos < *aγri-* *anθrop(os)*
"wild man" < wild man
vs.
***aγri**o* *`anθropos*

unless we deal with a coordinative relation between the members of a compound, as in the example

*pijen**o**`erx(ome)* < *pijen-* *erx(ome)*

"(I) come (and) go" < go come

vs.

*****pije`nerx(ome)*

In PC-KIMMO, the linking-vowel phenomenon is handled by an epenthesis two-level rule, which succeeds only in capturing the case where the second member of the compound construction begins by a consonant. It fails to account for the case where the rule is activated in presence of a coordinative relation between the two members of the construction.

C. A proposal for a modification of the formalism

We have already shown that a unification-based word grammar, consisting of a set of context-free rules, and augmented by feature structures, such as the one used by the PC-KIMMO system, is sufficient for a unified treatment of inflection, and compounding, without taking into consideration word-internal phonological change, syllabification and stress. In order to handle these phonological phenomena in a principled way, we would like to propose that the grammar should also be augmented with an activation of phonological rules.

Thus, the word grammar should consist of three components:

- A set of context-free rules
- A set of constraints on the feature structures
- A mechanism of selective activation of phonological rules

Note, that lexical phonological rules are formulated as two-level rules, applying at morpheme boundaries. Yet, there is no need for a massive parallel activation of these rules since in cases, such as the one reported in 3.2, their activation is triggered by information that is present on the feature structure. Moreover, we deal with stress phenomena at the level of the overall word structure. Thus, we avoid their representation at a segmental level, which is rather awkward from the point of view of linguistic knowledge-representation.

IV. CONCLUSION

In this paper, we discussed some problems in representing and processing some morpho-phonological phenomena in a natural language, like Greek. After discussing certain drawbacks of the two-level model, we proposed to modify the word grammar and the application of two-level rules, in order to reach a unified treatment of inflection, compounding, and word-internal phonological changes. In particular, we decided to augment the word grammar by a mechanism of selective application of two-level rules, applying at morpheme boundaries, and taking into consideration feature marking and structural constraints. In this way, stress and syllabification, which are indispensable for an efficient morphological analysis of Greek compound constructions, are handled in a satisfactory way.

REFERENCES

[1] E.L. Antworth, *PC-KIMMO: A Two-Level Processor for Morphological Analysis [Occasional Publications in Academic Computing 16],* Summer Institute of Linguistics, Dallas TX,1990.

[2] E.L. Antworth, "Morphological parsing with a unification-based word grammar", in *Proceedings of the North-Texas Natural Language Workshop*, University of Texas, Arlington,1994.

[3] E. Galiotou and A. Ralli, "Parsing Deficiencies of the PC-KIMMO System", in *Proceedings of the 2nd Hellenic Conference on Artificial Intelligence (Companion Volume)*, Aristotle University, Thessaloniki, 2002, pp.53-64.

[4] R. Kaplan and M. Kay, "Phonological rules and finite state transducers", in *Proceedings of the ACL/LSA Conference*, NewYork, 1981.

[5] L. Karttunen, "KIMMO: A general morphological processor*", Texas Linguistic Forum, 22*, 1983, pp. 163-186.

[6] L. Karttunen. and K. Oflazer (eds.), Special Issue on Finite State Methods in NLP, *Computational Linguistics, 26*, 1, 2000.

[7] Koskenniemi, K, "Two-level model for morphological analysis", in *Proceedings of IJCAI'83*, 1983, pp. 683-68.

[8] K. Koskenniemi, *Two-level Morphology: A General computational Model for Word-Form Recognition and Production*, Ph.D. Thesis, University of Helsinki, 1983.

[9] M. Nespor and A. Ralli, "Morphology-Phonology interface: phonological domains in Greek compounds", *in The Linguistic Review, 13*, 1996, pp. 357-382.

[10] A. Ralli, "Compounds in Modern Greek", in *Rivista di Linguistica, 4*, 1, 1992, pp.143-174.

[11] 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.

[12] A. Ralli, "A feature-based analysis of Greek nominal inflection", *Glossolojia, 11-12*, 2000, pp.201-227.

[13] A. Ralli, " The role of morphology in gender determination: evidence from Modern Greek" , in *Linguistics 40,* 3, pp. 519-551.

[14] A. Ralli, Μορφολογία (Morphology), Patakis, Athens, forthcoming.

[15] A. Ralli and E. Galiotou, "A prototype for a computational analysis of Modern Greek compounds", Asymmetry Conference, Université de Québec, à Montréal, May 2001.

[16] K. Sgarbas, N. Fakotakis and G. Kokkinakis, "A PC-KIMMO based morphological analysis of Modern Greek", in *Literary and Linguistic Computing, 10*, 3, 1995, pp. 189-201.

[17] S.M. Shieber, *An Introduction to Unification-Based Approaches to Grammar [CSLI Lecture Notes No 4],* Stanford, CA, 1986.

[18] R. Sproa t, *Morphology and Computation*, MIT Press, 1992.

[19] L. Touratzidis and A. Ralli, "Stress in Greek inflected forms: A computational treatment" , in *Language and Speech*, 35, 1992, pp. 435-453.

[20] K. Wallace, K. (ed.), *Morphology as a Computational Problem, UCLA Occasional Papers #7*, Dept. of Linguistics, UCLA, 1998.