Computing Prosodic Morphology

George Anton Kiraz[*]
University of Cambridge (St John's College)
Computer Laboratory
Pembroke Street
Cambridge CB2 1TP
George.Kiraz@cl.cam.ac.uk

Abstract

This paper establishes a framework under which various aspects of prosodic morphology, such as templatic morphology and infixation, can be handled under two-level theory using an implemented multi-tape two-level model. The paper provides a new computational analysis of root-and-pattern morphology based on prosody.

1 Introduction

Prosodic Morphology (McCarthy and Prince, 1986, et seq.) provides adequate means for describing non-linear phenomena such as infixation, reduplication and templatic morphology. Standard two-level systems proved to be cumbersome in describing such operations – see (Sproat, 1992, p. 159 ff.) for a discussion. Multi-tape two-level morphology (Kay, 1987; Kiraz, 1994, et. seq.) addresses various issues in the domain of non-linear morphology: It has been used in analysing root-and-pattern morphology (Kiraz, 1994), the Arabic broken plural phenomenon (Kiraz, 1996a), and error detection in non-concatenative strings (Bowden and Kiraz, 1995). The purpose of this paper is to demonstrate how non-linear operations which are motivated by prosody can also be described within this framework, drawing examples from Arabic.

The analysis of Arabic presented here differs from earlier computational accounts in that it employs new linguistic descriptions of Arabic morphology, viz. moraic and affixational theories (McCarthy and Prince, 1990b; McCarthy, 1993). The former argues that a different vocabulary is needed to represent the pattern morpheme according to the Prosodic Morphology Hypothesis (see §1.1), contrary to the earlier CV model where templates are represented as sequences of Cs (consonants) and Vs (vowels). The latter departed radically from the notion of root-and-pattern morphology in the description of the Arabic verbal stem (see §3).

The choice of the linguistic model depends on the application in question and is left for the grammarian. The purpose here is to demonstrate that multi-tape two-level morphology is adequate for representing these various linguistic models.

The following convention has been adopted. Morphemes are represented in braces, { }, and surface forms in solidi, / /. In listings of grammars and lexica, variables begin with a capital letter.

The structure of the paper is as follows: Section 2 demonstrates how Arabic templatic morphology can be analysed by prosodic terms, and section 3 looks into infixation; finally, section 4 provides some concluding remarks. The rest of this section introduces prosodic morphology and establishes the computational framework behind this presentation.

1.1 Prosodic Morphology

There are three essential principles in prosodic morphology (McCarthy and Prince, 1990a; McCarthy and Prince, 1993). They are:

(1) a. PROSODIC MORPHOLOGY HYPOTHESIS. Templates are defined in terms of the authentic units of prosody: mora (μ), syllable (σ), foot (Ft), prosodic word (PrWd).

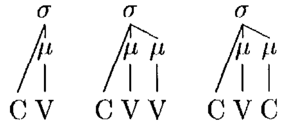
b. TEMPLATE SATISFACTION CONDITION. Satisfaction of templates constraints is obligatory and is determined by the principles of prosody, both universal and language-specific.

[*] Supported by a Benefactor Studentship from St John's College. This research was done under the supervision of Dr Stephen G. Pulman.

c. Prosodic Circumscription. The domain to which morphological operations apply may be circumscribed by prosodic criteria as well as by the more familiar morphological ones.

In the **Prosodic Morphology Hypothesis**, mora is the unit of syllabic weight; a monomoraic syllable, σ_{μ}, is light (L), and a bimoraic syllable, $\sigma_{\mu\mu}$, is heavy (H). The most common types of syllables are: open light, CV, open heavy, CVV, and closed heavy, CVC. This typology is represented graphically in (2).

(2)



Association of Cs and Vs to templates is based on the **Template Satisfaction Condition**. Association takes the following form: a node σ always takes a C, and a mora μ takes a V; however, in bimoraic syllables, the second μ may be associated to either a C or a V.[1]

Prosodic Circumscription (PC) defines the domain of morphological operations. Normally, the domain of a typical morphological operation is a grammatical category (root, stem or word), resulting in prefixation or suffixation. Under PC, however, the domain of a morphological operation is a prosodically-delimited substring within a grammatical category, often resulting in some sort of infixation. The essential for PC is a parsing function Φ of the form in (3).

(3) Parsing Function
Φ(C, E)

Let B be a base (i.e. stem or word). The function Φ returns the constituent C that sits on the edge E $\in$ {`right`, `left`} of the base B. The result is a factoring of B into: **kernel**, designated by B:Φ, which is the string returned by the parsing function, and **residue**, designated by B/Φ, which is the remainder of B. The relation between B:Φ and B/Φ is given in (4), where $\frown$ is the concatenation operator.

(4) Factoring of B by Φ
$B = B{:}\Phi \frown B/\Phi$

To illustrate this, let B = /katab/; applying the function $\Phi(\sigma_{\mu}$, Left) on B factors it into: (i) the kernel B:Φ = /ka/, and (ii) the residue B/Φ = /tab/.

A morphological operation O (e.g. O = "Prefix {t}") defined on a base B is denoted by O(B). There are two types of PC: **positive** (PPC) and **negative** (NPC). In PPC, the domain of the operation is the kernel B:Φ; this type is denoted by O:Φ and is defined in (5a). In NPC, the domain is the residue B/Φ; this type is denoted by O/Φ and is defined in (5b).

(5) Definition of PPC and NPC
a. PPC, $O{:}\Phi(B) = O(B{:}\Phi) \frown B/\Phi$
b. NPC, $O/\Phi(B) = B{:}\Phi \frown O(B/\Phi)$

In other words, in PPC, O applies to the kernel B:Φ, concatenating the result with the residue B/Φ; in NPC, O applies to the residue B/Φ, concatenating the result with the kernel B:Φ. Examples are provided in section 3.

1.2 Multi-Tape Two-Level Formalism

Two-level morphology (Koskenniemi, 1983) defines two levels of strings in recognition and synthesis: lexical strings represent morphemes, and surface strings represent surface forms. Two-level rules map the two strings; the rules are compiled into finite state transducers, where lexical strings sit on one tape of the transducers and surface strings on the other.

Multi-tape two-level morphology is an extension to standard two-level morphology, where more than one lexical tape is allowed. The notion of using multiple tapes first appeared in (Kay, 1987). Motivated by Kay's work, (Kiraz, 1994) proposed a multi-tape two-level model. The model adopts the formalism in (6) as reported by (Pulman and Hepple, 1993).

$$(6)\quad \begin{array}{lllllll} \text{LLC} & - & \text{Lex} & - & \text{RLC} & \{\Rightarrow, \Leftrightarrow\} \\ \text{LSC} & - & \text{Surf} & - & \text{RSC} & \end{array}$$

where LLC is the left lexical context, Lex is the lexical form, RLC is the right lexical context, LSC is the left surface context, Surf is the surface form, and RSC is the right surface context.

The special symbol * indicates an empty context, which is always satisfied. The operator $\Rightarrow$ states that Lex *may* surface as Surf in the given context, while the operator $\Leftrightarrow$ adds the condition that when Lex appears in the given context, then the surface description *must* satisfy Surf. The latter caters for obligatory rules. A lexical string maps to a surface string iff (1) they can be partitioned into pairs of lexical-surface subsequences, where each pair is licenced by a rule, and (2) no partition violates an obligatory rule.

[1] Other conventions associate consonant melodies left-to-right to the moraic nodes, followed by associating vowel melodies to syllable-initial morae.

One of the extensions introduced in the multitape version is that all expressions in the lexical side of the rules (i.e. LLC, LEX and RLC) are n-tuple of regular expressions of the form (x_1, x_2, ..., x_n). The ith expression refers to symbols on the ith tape. When $n = 1$, the parentheses can be ignored; hence, (x) and x are equivalent.[2]

2 Templatic Morphology

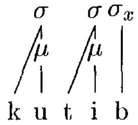
Templatic morphology is best exemplified in Semitic root-and-pattern morphology. This section sets a framework under which templatic morphology can be described using (augmented) two-level theory. Our presentation differs from previous proposals[3] in that it employs prosodic morphology in the analysis of Arabic, rather than earlier CV accounts. Arabic verbal forms appear in (7) in the passive (rare forms are not included).

(7) ARABIC VERBAL MEASURES (1-8, 10)

1	kutib	6	tukuutib
2	kuttib	7	nkutib
3	kuutib	8	ktutib
4	ʔuktib		
5	tukuttib	10	stuktib

(McCarthy, 1993) points out that Arabic verbal forms are derived from the base template in (8), which represents Measure 1. σ_x represents an extrametrical consonant; that is, the last consonant in a stem.

(8) ARABIC BASE TEMPLATE



The remaining measures are derived from the base template by affixation; they have no templates of their own. The simplest operation is prefixation, e.g. {n} + Measure 1 → /nkutib/ (Measure 7). Measures 4 and 10 are derived in a similar fashion, but undergo a rule of syncope as shown in (9).

(9) DERIVATION OF MEASURES 4 AND 10

Syncope: V $\longrightarrow \phi$ /[CVC ___ CVC]$_{\text{stem}}$

a. Measure 4: ʔu + kutib $\longrightarrow$ */ʔukutib/ $\xrightarrow{syncope}$ /ʔuktib/

b. Measure 10: stu + kutib $\longrightarrow$ */stukutib/ $\xrightarrow{syncope}$ /stuktib/

The following lexicon and two-level grammar demonstrate how the above measures can be analysed under two-level theory. The lexicon maintains four tapes: pattern, root, vocalism and affix tapes.

1	$\{\sigma_\mu\sigma_\mu\sigma_x\}$	`pattern:[measure=(1-8,10)]`
2	{ktb}	`root:[measure=(1-4,6-8,10)]`

3	{ui}	`vocalism:[tense=perf,`
		`voice=pass]`

4	{ʔV}	`verb_affix:[measure=4]`
4	{n}	`verb_affix:[measure=7]`
4	{stV}	`verb_affix:[measure=10]`

The first column indicates the tape on which the morpheme sits, and the second column gives the morpheme. Each lexical entry is associated with a category and a feature structure of the form `cat:FS` (column 3). Feature values in parentheses are disjunctive and are implemented using boolean vectors (Mellish, 1988; Pulman, 1994).

$\{\sigma_\mu\sigma_\mu\sigma_x\}$ is the base-template. {ktb} 'notion of writing' is the root; it may occur in all measures apart from Measure 5.[4] {ui} is the perfective passive vocalism. The remaining morphemes represent the affixes for Measures 4, 7 and 10. Notice that the vowel in the affixes of Measures 4 and 10 is a variable V. This makes it possible for the affix to have a different vowel according to the mood of the following stem, e.g. [a] in /ʔaktab/ (Measure 4, active) and [u] in /ʔuktib/ (Measure 4, passive).

Since the lexicon declares 4 lexical tapes, each lexical expression in the two-level grammar must be at most a 4-tuple. A grammar for the derivation of the cited data appears below.

R1	*	–	$\langle\sigma_\mu$,C,V,$\varepsilon\rangle$	–	*	⇒
	*	–	CV	–	*	
R2	*	–	$\langle\sigma_x$,C,ε,$\varepsilon\rangle$	–	$\langle$+,+,+,$\varepsilon\rangle$	⇔
	*	–	C	–	*	
R3	*	–	$\langle\varepsilon,\varepsilon,\varepsilon$,A$\rangle$	–	*	⇒
	*	–	A	–	*	
R4	$\langle$X,*,ε,$\varepsilon\rangle$	–	$\langle$+,+,+,$\varepsilon\rangle$	–	*	⇒
	*	–	ε	–	*	
R5	$\langle\varepsilon,\varepsilon,\varepsilon$,A$\rangle$	–	$\langle\varepsilon,\varepsilon,\varepsilon$,+$\rangle$	–	*	⇒
	*	–	ε	–	*	

[2] Our implementation interprets rules directly (see (Kiraz, 1996c)); hence, we allow unequal representation of strings. If the rules were to be compiled into automata, a genuine symbol, e.g. 0, must be introduced by the rule compiler. For the compilation of our formalism into automata, see (Grimley-Evans et al., 1996).

[3] Non-linear proposals include (Kay, 1987), (Kornai, 1991), (Wiebe, 1992), (Narayanan and Hashem, 1993), (Bird and Ellison, 1994) and (Kiraz, 1994). A working system for Arabic is reported by (Beesley et al., 1989; Beesley, 1990; Beesley, 1991).

[4] Roots do not occur in all measures in the literature. Each root is lexically marked with the measures it occurs in.

$$\text{R6}\quad \begin{array}{cccccc} * & - & \langle\sigma_\mu,\text{C},\text{V},\varepsilon\rangle & - & * & \Leftrightarrow \\ \text{C}_1\text{V} & - & \text{C} & - & \text{C}_2\text{V}_1\text{C}_3 & \end{array}$$

where C_i=`radical`, V_i=`vowel`, A=`verbal affix`, and $X \neq +$.

Rule R1 handles monomoraic syllables mapping $(\sigma_\mu,C,V,\varepsilon)$ on the lexical tapes to CV on the surface tape. Rule R2 maps the extrametrical consonant in a stem (i.e. the last consonant in a stem) to the surface. Rule R3 maps an affix symbol from the fourth tape to the surface. Rules R4 and R5 delete the boundary symbols from stems and affixes, respectively. Finally, rule R6 simulates the syncope rule in (9); note that V in LSC must unify with V in LEX, ensuring that the vowel of the affix has the same quality as that of the stem, e.g. /ʔaktab/ and /ʔu+ktib/ (Measure 4).

The two-level analysis of the cited forms appears below – ST = surface tape, PT = pattern tape, RT = root tape, VT = vocalism tape, and AT = affix tape.

Measure 1

				AT
u	i		+	*VT*
k	t	b	+	*RT*
σ_μ	σ_μ	σ_x	+	*PT*
1	1	2	4	
ku	ti	b		*ST*

Measure 4

ʔ	u	+					
			u	i		+	
			k	t	b	+	
			σ_μ	σ_μ	σ_x	+	
3	3	5	6	1	2	4	
ʔ	u		k	ti	b		

Measure 7

n	+					*AT*
		u	i		+	*VT*
		k	t	b	+	*RT*
		σ_μ	σ_μ	σ_x	+	*PT*
3	5	1	1	2	4	
n		ku	ti	b		*ST*

Measure 10

s	t	u	+				
				u	i		+
				k	t	b	+
				σ_μ	σ_μ	σ_x	+
3	3	3	5	6	1	2	4
s	t	u		k	ti	b	

The numbers between the two levels indicate the rule numbers in (8) which sanction the sequences. The remaining Measures involve infixation and are discussed in the next section.

3 Infixation

Standard two-levels models can describe some classes of infixation, but resorting to the use of *ad hoc* diacritics which have no linguistic significance, e.g. (Antworth, 1990, p. 156). This section presents a framework for describing infixation rules using our multi-tape two-level formalism. This is illustrated here by analysing Measures 2 and 8 of the Arabic verb. Measure 2, /kuttib/, is derived by prefixing a mora to the base template under NPC. The operation is O = 'prefix μ' and the rule is $O/\Phi(\sigma_\mu$, Left). The new mora is filled by the spreading of the adjacent (second) consonant. The steps of the derivation are:

```
O/Φ(kutib) = kutib:Φ * O(kutib:Φ)
           = ku * O(tib)
           = ku * μtib
           = ku * ttib
           = kuttib
```

Measure 8, /ktutib/, is derived by the affixation of a {t} to the base template under NPC. The operation is O = 'prefix {t}'; the rule is O/Φ(C, Left), where C is a consonant. The process is:

```
O/Φ(kutib) = kutib:Φ * O(kutib:Φ)
           = k * O(utib)
           = k * tutib
           = ktutib
```

The following two-level grammar builds on the one discussed in section 2. The following lexical entry gives the Measure 8 morphemes.

4 {t} `verb_affix:[measure=8]`

The additional two-level rules are:

$$\text{R7}\quad \begin{array}{cccccc} \langle\sigma_\mu,\text{C}_1,\text{V}_1,\varepsilon\rangle & - & \varepsilon & - & \langle\sigma_\mu,\text{C},*,\varepsilon\rangle & \Rightarrow \\ * & - & \text{C} & - & * & \end{array}$$

Features: `[measure=(2,5)]`

$$\text{R8}\quad \begin{array}{cccccc} * & - & \langle\sigma_\mu,\text{C},\text{V},\text{A}\rangle & - & * & \Rightarrow \\ * & - & \text{CAV} & - & * & \end{array}$$

Features: `[measure=8]`

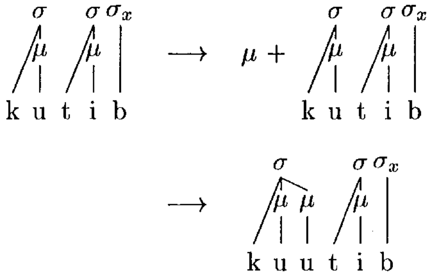
where C_i=`radical`, V_i=`vowel`, A=`verbal affix`, and $X \neq +$.

Rules R7-R8 are measure-specific. Each rule is associated with a feature structure which must unify with the feature structures of the affected lexical entries. This ensures that each rule is applied only to the proper measure.

R7 handles Measure 2; it represents the operation O = 'prefix μ' and the rule $O/\Phi(\sigma_\mu$, Left) by placing B:Φ in LLC and the residue B/Φ in RLC, and inserting a consonant C (representing μ) on the surface. The filling of μ by the spreading of the second radical is achieved by the unification of C in LEX with C in RLC.

R8 takes care of Measure 8; it represents the operation O = 'prefix {t}' and the rule O/Φ(C, Left). Note that one cannot place B:Φ and B/Φ in LLC and RLC, respectively, as the case in R7 because the parsing function cuts into the first syllable.

One remaining Measure has not been discussed, Measure 3. It is derived by prefixing the base template with μ. The process is as follows:



σ σ σ_x
μ μ
k u t i b
μ +
σ σ σ_x
μ μ
k u t i b
σ σ σ_x
μ μ μ
k u u t i b

The corresponding two-level rule follows. It adds a μ by lengthening the vowel V into VV.

	*	–	$\langle\sigma_\mu$,C,V,$\varepsilon\rangle$	–	*	$\Rightarrow$
R9	*	–	CVV	–	*	

Features: `[measure=(3,6)]`

The two-level derivations are:

						AT
	u	i		+		*VT*
Measure 2	k	t		b	+	*RT*
	σ_μ	σ_μ		σ_x	+	*PT*
	1	7	1	2	4	
	ku	t	ti	b		*ST*

					AT
	u	i		+	*VT*
Measure 3	k	t	b	+	*RT*
	σ_μ	σ_μ	σ_x	+	*PT*
	9	1	2	4	
	kuu	ti	b		*ST*

	t	+				*AT*
	u	i		+		*VT*
Measure 8	k	t		b	+	*RT*
	σ_μ	σ_μ		σ_x	+	*PT*
	8	5	1	2	4	
	ktu		ti	b		*ST*

Finally, Measures 5 and 6 are derived by prefixing {tu} to Measures 2 and 3, respectively.

4 Conclusion

This paper have demonstrated that multi-tape two-level systems offer a richer and more powerful devices than those in standard two-level models. This makes the multi-tape version capable of modelling non-linear operations such as infixation and templatic morphology.

The rules and lexica samples reproduced here are based on a larger morphological grammar written for the SemHe implementation (a multi-tape two-level system) – for a full description of the system, see (Kiraz, 1996c; Kiraz, 1996b).

References

Antworth, E. (1990). *PC-KIMMO: A two-Level Processor for Morphological Analysis.* Occasional Publications in Academic Computing 16. Summer Institute of Linguistics, Dallas.

Beesley, K. (1990). Finite-state description of Arabic morphology. In *Proceedings of the Second Cambridge Conference: Bilingual Computing in Arabic and English.*

Beesley, K. (1991). Computer analysis of Arabic morphology. In Comrie, B. and Eid, M., editors, *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics.* Benjamins, Amsterdam.

Beesley, K., Buckwalter, T., and Newton, S. (1989). Two-level finite-state analysis of Arabic morphology. In *Proceedings of the Seminar on Bilingual Computing in Arabic and English.* The Literary and Linguistic Computing Centre, Cambridge.

Bird, S. and Ellison, T. (1994). One-level phonology. *Computational Linguistics*, 20(1):55–90.

Bowden, T. and Kiraz, G. (1995). A morphographemic model for error correction in nonconcatenative strings. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 24–30.

Grimley-Evans, E., Kiraz, G., and Pulman, S. (1996). Compiling a partition-based two-level formalism. In *COLING-96.*

Kay, M. (1987). Nonconcatenative finite-state morphology. In *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, pages 2–10.

Kiraz, G. (1994). Multi-tape two-level morphology: a case study in Semitic non-linear morphology. In *COLING-94: Papers Presented to the 15th International Conference on Computational Linguistics*, volume 1, pages 180–6.

Kiraz, G. (1996a). Analysis of the Arabic broken plural and diminutive. In *Proceedings of the 5th International Conference and Exhibition on Multi-Lingual Computing.* Cambridge.

Kiraz, G. (1996b). *Computational Approach to Non-Linear Morphology.* PhD thesis, University of Cambridge.

Kiraz, G. (1996c). ȘEMḤE: A generalised two-level system. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.*

Kornai, A. (1991). *Formal Phonology.* PhD thesis, Stanford University.

Koskenniemi, K. (1983). *Two-Level Morphology.* PhD thesis, University of Helsinki.

McCarthy, J. (1993). Template form in prosodic morphology. In Stvan, L. *et al.*, editor, *Papers from the Third Annual Formal Linguistics Society of Midamerica Conference*, pages 187–218. Indiana University Linguistics Club, Bloomington.

McCarthy, J. and Prince, A. (1986). Prosodic morphology. ms.

McCarthy, J. and Prince, A. (1990a). Foot and word in prosodic morphology: The Arabic broken plural. *Natural Language and Linguistic Theory*, 8:209–83.

McCarthy, J. and Prince, A. (1990b). Prosodic morphology and templatic morphology. In Eid, M. and McCarthy, J., editors, *Perspectives on Arabic Linguistics II: Papers from the Second Annual Symposium on Arabic Linguistics*, pages 1–54. Benjamins, Amsterdam.

McCarthy, J. and Prince, A. (1993). Prosodic morphology. ms.

Mellish, C. (1988). Implementing systemic classification by unification. *Computational Linguistics*, 14(1):40–51.

Narayanan, A. and Hashem, L. (1993). On abstract finite-state morphology. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–304.

Pulman, S. (1994). Expressivity of lean formalisms. In Markantonatou, S. and Sadler, L., editors, *Grammatical Formalisms: Issues in Migration*, Studies in Machine Translation and Natural Language Processing: 4, pages 35–59. Commission of the European Communities.

Pulman, S. and Hepple, M. (1993). A feature-based formalism for two-level phonology: a description and implementation. *Computer Speech and Language*, 7:333–58.

Sproat, R. (1992). *Morphology and Computation.* MIT Press, Cambridge Mass.

Wiebe, B. (1992). Modelling autosegmental phonology with multi-tape finite state transducers. Master's thesis, Simon Fraser University.