Computational Nonlinear Morphology with Emphasis on Semitic Languages

George Anton Kiraz
(Beth Mardutho: The Syriac Institute)

Cambridge: Cambridge University Press (Studies in natural language processing, edited by Branimir Boguraev and Steven Bird), 2001, xxi+171 pp; hardbound, ISBN 0-521-63196-3, $60.00

Reviewed by
Markus Walther
Panasonic Speech Technology Laboratory

1. Introduction

Computational morphology would be an almost trivial exercise if every language were like English. Here, chopp-ing off the occasion-al affix-es, of which there are not too many, is sufficient to isolate the stem, perhaps modulo a few (morpho)graphemic rules to handle phenomena like the consonant doubling we just saw in *chopping*. This relative ease with which one can identify the core meaning component of a word explains the success of rather simple stemming algorithms for English or the way in which most part-of-speech (POS) taggers get away with just examining bounded initial and final substrings of unknown words for guessing their parts of speech. In contrast, this book outlines a computational approach to morphology that explicitly includes languages from the Semitic family, in particular Arabic and Syriac, where the linearity hypothesis—every word can be built via string concatenation of its component morphemes—seems to break down (we will take up the validity of that assumption below).

Example 1 illustrates the problem at hand with Syriac verb forms of the root $\{q_1ṭ_2l_3\}$ 'notion of killing' (from Kiraz [1996]).

(1)

	Stem shape	*Form*	*Morphs*	*Gloss*
a.	$C_1C_2V_2C_3$	*qṭal*	a_1a_2 *past act.*	*he killed*
b.		*neqṭol*	*ne-* 3 *sg. m.*, a_1o_2 *fut.*	*he will kill*
c.		ܐܶ*eθqṭel*	ܐ *eθ- refl.*, a_1e_2 *past pass.*	*he was killed*
d.	$C_1V_1C_2C_3$	*qaṭleh*	a_1a_2 *past act.*, *-eh obj.*	*he killed-OBJ*
e.	$C_1C_2C_3$	*neqṭlu*ː *n*	*ne- -u* ː *n 3 pl. m.*, a_1o_2 *fut.*	*they (m.) will kill*

Notice the use of subscripts as a visual aid in pairing up abstract consonantal (C) and vocalic (V) stem positions with concrete segments. The stem shapes show how root and tense/aspect morphemes are interdigitated. Also evident is the considerable variability in stem vowel (non)realization, leading to vowelless stems in the extreme case (1e).

2. Content

The opening chapter begins by specifying the intended wide audience, namely computational, theoretical, and applied linguists as well as Semitists. It then addresses linguistic preliminaries, including brief introductions to morphology and autosegmental phonology, before proceeding to some formal language theory and unification. Introductory applications of these to selected morphology and phonology problems are given. Very briefly, the bare basics of Semitic noun and verb morphology are touched upon as well as some peculiarities of its predominant writing system.

Chapter 2 is a very useful survey of three mainstream approaches to the formal description of Semitic word formation that differ in terms of which units form the template, that is, the central sequencing device (CV vs. moraic), and how many templates are assumed (affixational approach). Here Kiraz strictly focuses on pre-optimality-theoretic work by John McCarthy and Alan Prince, two influential theorists in generative linguistics. Notably the author also draws attention to aspects of Semitic morphology beyond the stem, highlighting the existence of various affixation processes as well as phonological effects such as vowel deletions sensitive to syllable structure.

Chapter 3 begins by mentioning the work of Kaplan and Kay (1994) on cascaded finite-state rules but mostly focuses on further developments of the two-level model (Koskenniemi 1983) for parallel rule application in a finite-state setting, since Kiraz intends to use an extended formalism from that class. Among the modifications reviewed are mapping of sequences rather than single symbols only, unequal-length mappings, unification over finite-valued features, and proper treatment of obligatory rules.

Chapter 4 prepares the ground for Kiraz's own work by reviewing no less than nine different approaches to Semitic computational morphology. They broadly fall into two classes, one following the autosegmental, multitiered approach, whether expressed by mappings between several automaton tapes or intensional descriptions that codescribe a single tape. The other class follows no particular theory but often uses regular set intersection to combine root, template, and vowel pattern.

The central Chapter 5 finally introduces Kiraz's own multitier formalism. Here we find comprehensive descriptions and formal definitions of the lexicon and rewrite-rule components. The former consists of sublexica corresponding to the various lexical tiers or tapes, whereas the latter allows two-level-style context restriction and surface coercion rules. All the modifications discussed in chapter 3 are incorporated here, and proposals for handling morphotactics are described as well.

Chapter 6 now applies the multitier formalism to selected problems of Arabic morphology. It details the three approaches of chapter 2 to verb stem formation, giving formal rules and lexicon entries that allow the reader to simulate sample stem derivations in Kiraz's framework. With regard to noun morphology, "broken" plurals like *xaatam* 'signet-ring (sg.)' $\sim$ *xawaatim* '(pl.)' receive a formal analysis as well. Kiraz discusses issues of nonlinearity versus linearity and generation of partially voweled spellings before finishing the chapter with a rule-based treatment of glyph alternations in Syriac script.

Chapter 7 develops the compilation of Kiraz's formalism into multitape automata, broadly using the concepts and methodology of Kaplan and Kay while introducing additional regular operators for n-way regular relations. Because the different stages can get quite involved technically, they are illustrated step by step with the help of simple examples and automaton diagrams.

The book concludes in Chapter 8 by first presenting a short discussion of applications of the formalism to general autosegmental problems, illustrated with an

African tone language. Then it touches on the subjects of disambiguation of Semitic orthographic representations (high ambiguity due to absence of short vowels), semantics in Semitic (sense disambiguation), and productivity (mainly extension of existing roots to previously unused patterns). Interestingly, Kiraz speculates that addressing productivity might involve weighted automata to express the preference for roots to attach to lexically known patterns without completely ruling out a new-word interpretation.

Finally, five pages of references and three indices are provided. The book appears to be carefully edited, has a professional layout, and is remarkably free of typographic and spelling errors.

3. Critique

The author stresses (p. xv) that the research for this book, originally his Ph.D. thesis, took place between 1992 and 1996. With five years to publication, there is considerable risk of new developments in the field (or a revival of old ideas) that could provide competing insight or weaken central claims. This section will discuss some of the more problematic aspects of this book in this regard.

But first, what about its suitability for the stated target audience? Although bridging the gap between the separate disciplines that share an interest in the subject is certainly a laudable goal, this reviewer is quite unsure whether the book succeeds in meeting it. The Semitist will probably feel overwhelmed by the amount of mathematical formalism, without getting rewarded in the end by, say, application to interesting comparative or diachronic problems from his field of interest. Theoretical linguists will in addition recognize immediately that the book does not cover constraint-based approaches like optimality theory, which from the very beginning were strongly motivated by prosodic morphology (McCarthy and Prince 1993), of which the Semitic kind is a fine example. If merely adapting now-abandoned analyses to a computational setting is not a particularly strong selling point for this group, then neither is the absence of a detailed treatment of some of the more interesting issues that Semitic presents, such as how to capture its morphological richness with few parameterized or prioritized principles, how to regularize the apparent irregularity of weak verbs, and so on. That leaves the computational linguist who wants to, say, build a practical morphological analyzer for Arabic or understand the minimal computational requirements for a plausible model of morphology that includes Semitic languages.

Following the recent trend toward data-intensive, empirically oriented computational linguistics, such a reader will probably first want to see a decent introduction to the phenomena at hand. But what they get is rather disappointing. Kiraz does describe the Arabic "broken" plural, giving a number of example pairs, but without proper discussion of its productivity and the corresponding "sound" plural it is a bit hard to understand why it is worth being modeled by rules instead of lexical listing. For verbs, no exemplary paradigms of surface forms are given at all, and no tables list nontrivial excerpts of the morphological system of a language as unfamiliar as Syriac. When Arabic stems are presented (page 34), the reader has to wait 28 pages to be informed that, actually, the form */nkutib/* is pronounced [ʔinkutib] 'write (measure 7, pass.)'. Of course, this makes a huge difference: The former is prosodically ill formed, unlike the latter, whose prefix *ʔin-* is a well-formed syllable. Insightful linguistic analysis is hardly possible when using defective data, yet Kiraz bases his formal analysis on them (page 104f). Regrettably we are often not given enough detail about the prosodic

systems of both languages: Avoidance of initial CC clusters in Arabic is mentioned in passing, but is it exceptionless? And what about the distribution of the same in Syriac, where such clusters are allowed? In a section on neologisms (page 152), only the expert will not be puzzled when Kiraz cites two such forms without glosses; one cannot even pronounce the Syriac form of the two because the transliterated vowel symbol å is not explained (page xx). In sum, the nonspecialist is given too little of the big picture to be able to come up with alternative ideas about plausible models for the data.

Next, the reader may start wondering whether it is actually true that "[u]sing the nonlinear model is *crucial* for developing Semitic systems" (page 110, emphasis added). Kiraz himself never questions the tradition that interprets the conceptual autonomy of consonantal root, template, and vowel sequence[1] as technical nonlinearity.

He does show, however, that actually a nonlinear representation is harmful everywhere but in the stem, for example, leading to duplication of rules when coverage is extended to affixed forms (page 112f). As a consequence, he must weaken his architecture to provide a second stage in which rules postprocess fully linearized verb stems; the same setup is proposed for broken plural formation in nouns, because vowel length and prosodic shape transfer from singular to plural and cannot be read off the components alone. A third, again linear, stage optionally deletes short vowels from the fully pronounceable surface form to map to partially voweled orthographic representations. At this point good scientific reductionism would seem to suggest trying to reduce nonlinearity to zero, but Kiraz offers no discussion of why any such alternative won't fly.

In fact, such an alternative has been proposed by Hudson (1986) for Arabic. In the briefest of sketches, a modernized version taken from Walther (1999) goes like this: We replace object strings by partial descriptions and encode stems with the help of optionality parentheses for zero-alternating vowels, for example, *q(a)t(o)l* for the future stem. While one such description denotes four surface strings, nonalternating affixes are represented without optional segments, giving *neq(a)t(o)l* after concatenation (cf. 1b). Using the central insight that the shape of entire word forms, not stems alone, is governed by syllable structure constraints, here (C)CV(V)(C), we are left with the set {*neqatol*,*neqtol*}. Assigning a weight to every realized vowel, we can finally apply a left-to-right greedy shortest-path algorithm to correctly prefer *neqtol* over *neqatol* because it omits an alternating stem vowel as early as possible. Note that left-to-right incrementality is psycholinguistically plausible and leads on average to an earlier recognition point for the root. This approach, which has been used to formulate sizeable morphological grammars for Tigrinya (Walther 1999) and Modern Hebrew (Walther 1998), can also be implemented in finite-state terms. With the aid of an inheritance-based formalism, redundancy in stem descriptions would be kept minimal, thus retaining a logical, but not object-level, autonomy of stem components while accommodating exceptions at the same time. In contrast to Kiraz's approach, which must employ baroque vowel deletion rules that operate right to left to edit the abstract stems under affixation, the constraint-based alternative sketched is much more explanatory in terms of why Semitic stems exhibit so much shape *variance* instead of the shape invariance predicted by Kiraz's rigid templates: They simply respond to both the language-particular restrictive syllable canon and universal demands for processing economy. Under this perspective, Semitic morphology is formally atemplatic

1 Although these are usually motivated both by descriptive economy and identifiable semantic contribution, Kiraz does not discuss the significant extent to which stems in Semitic languages like Modern Hebrew have noncompositional meanings that cannot be predicted from their components.

and concatenative, differing mainly by its regular use of vowel/zero alternation and ablaut (cf. sporadic cases like German *Segel* ∼ *Segl-er* 'sail ∼ sailor' and English *sing* ∼ *sang* ∼ *sung*).[2]

If its linguistic motivation is found wanting, perhaps the main strength of Kiraz's proposal comes from the technical side, with greater computational efficiency and just the right expressivity? In fact, this is what Kiraz seems to have in mind (pages 68, 111). When discussing related work that dispenses with multiple lexical tapes or tiers—while still sharing the template idea—he identifies intersection-based and mapping-based approaches as the main players. Simply put, in the former, consonantally underspecified template automata like *CaCaC* are intersected with vocalically underspecified root automata such as *kVtVb*, whereas in the latter, one rewrite rule is constructed per stem that specifies the linear arrangement of its components at compile time. In his critique Kiraz alleges that intersection loses bidirectionality; that is, parsing cannot reliably recover the root and the other components if given just stems, that one-rule-per-stem is highly redundant, and that both approaches are computationally expensive at compile time compared to his multitape approach.

Just as Kiraz modifies traditional automata, however, so can proponents of the intersection approach (which is similar to the alternative outlined above). For example, to recover whether a segment originates from root or vocalic pattern or affix, one could envision composite labels ⟨*segment*, *origin*⟩ on automata transitions, where *segment* parts match traditionally, whereas *origin*s are unioned together. The parse string would start out with empty *origin* sets.

As for the other advantage, compile-time efficiency: This is a notoriously risky argument, given that computers get faster all the time and that the main attractiveness of finite-state processing lies in its fast *run-time* behavior. In this regard it is curious that Kiraz cites his paper (Kiraz 2000) but does not incorporate the empirical evaluation from that paper into the book to strengthen his claim. In any case, recent results by Beesley and Karttunen (2000) show that fast compilation no longer necessitates a multitier model. Their proposal is that automata strings could themselves contain textual representations of regular expressions, with a new *compile-replace* operator allowing for in situ evaluation and substitution. This reduces compile time from hours to a few minutes for a large-scale Arabic morphology, using compile-replace for stem formation and composition with finite-state rule transducers to map to the surface forms. Although this might seem like an eclectic mix of different strategies, recall that Kiraz himself has a hybrid system with several composed stages: Does this imply that his multitier formalism by itself is not expressive enough for practical grammars?

To be sure, the book does have its strong sides, including good reviews of related work and an exposition of a particular multitape finite-state formalism that is detailed enough to allow the interested reader to implement it, and—if so desired—create a working morphology system for Semitic and other languages. Therefore I would recommend it as a useful source of inspiration for researchers in the field, as long as their foreseen applications are unaffected by the criticism presented above.

2 Kiraz (1996) defends the necessity of abstract stems such as Syriac **katab* because a rule that turns certain plosives into same-place fricatives applies after short vowels (→ **kaθav*), which may be deleted in surface forms (→ *kθav* 'he wrote'). In fact, however, a surface-true prosodic reformulation *can* account for his data: Those plosives fricativize after noncoda segments, here the complex-onset member *k* and nucleus *a*.

References

Beesley, Kenneth R. and Lauri Karttunen. 2000. Finite-state non-concatenative morphotactics. In Jason Eisner, Lauri Karttunen, and Alain Thériault, editors, *Proceedings of the Fifth Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-2000)*, pages 1–12, Luxembourg.

Hudson, Grover. 1986. Arabic root and pattern morphology without tiers. *Journal of Linguistics*, 22:85–122.

Kaplan, Ronald and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Kiraz, George Anton. 1996. Syriac morphology: From a linguistic model to a computational implementation. In R. Lavenant, editor, *VII Symposium Syriacum 1996*. Orientalia Christiana Analecta, Rome.

Kiraz, George Anton. 2000. Multitiered nonlinear morphology using multi-tape finite automata: A case study on Syriac and Arabic. *Computational Linguistics*, 26(1):77–105.

Koskenniemi, Kimmo. 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki, Helsinki.

McCarthy, John and Alan Prince. 1993. Prosodic morphology I: Constraint interaction and satisfaction. Technical Report RuCCS-TR-3, Rutgers University Center for Cognitive Science.

Walther, Markus. 1998. Computing declarative prosodic morphology. In Mark Ellison, editor, *Proceedings of the Fourth Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON 98)*, pages 11–20, Montreal.

Walther, Markus. 1999. *Deklarative prosodische Morphologie: Constraint-basierte Analysen und Computermodelle zum Finnischen und Tigrinya*. Niemeyer, Tübingen, Germany.

Markus Walther has published on computational phonology and morphology using logic-based and finite-state formalisms, with applications including reduplication and Semitic word formation in Tigrinya and Modern Hebrew. He now works in research and development for text-to-speech synthesis. Walther's address is Panasonic Speech Technology Laboratory, 3888 State Street #202, Santa Barbara, CA 93101; e-mail: mwalther@stl.research.panasonic.com; URL: www.markus-walther.de.