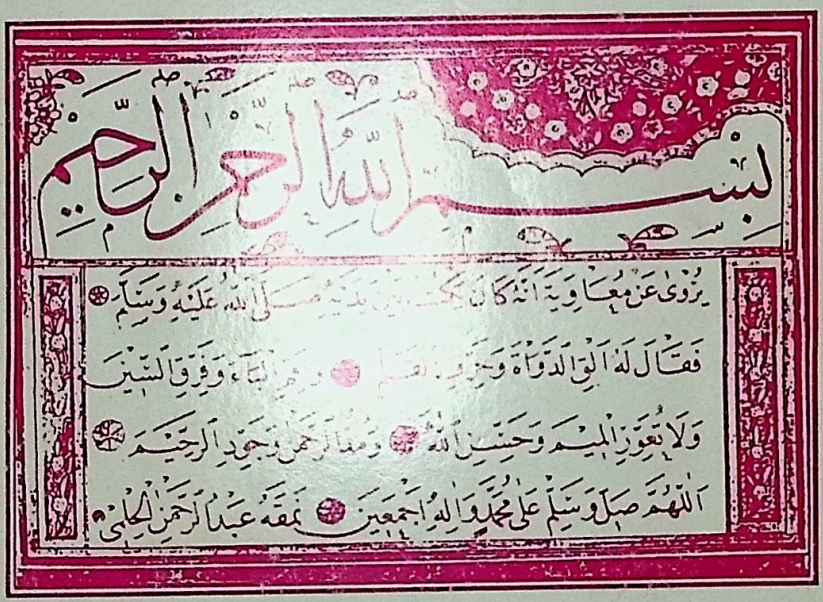
The Documentation Unit

The Centre for

Middle Eastern and Islamic Studies

University of Durham



Proceedings

of the 3rd International Conference and Exhibition on Multi-lingual Computing

(Arabic and Roman Script)

10-12 December 1992

Semitic Morphology:
A Linguistic and Computational Account

By George Anton Kiraz

Computer Laboratory
University of Cambridge (Peterhouse)
gk105@uk.ac.cam.phx

1. Introduction

This paper gives a brief account of the theoretical linguistic and computational backgrounds of Semitic computational morphology. The material is based on Kiraz (1992)*. The account presented here reflects studies which took place in the West.

This paper is organized as follows: Section 2 gives a description of the Semitic morpheme. Section 3 gives the linguistic theoretical background of Semitic morphology. Section 4 gives an account of computational morphology, and previous proposals for Semitic. Section 5 attempts to define a framework for Semitic morphology.

When citing Arabic (Arab.) or Syriac (Syr.) words, I use the following format: «Arab. *kataba*» 'he wrote', or «Syr. *ktab*» 'he wrote'. '√' indicates a root.

2. Morphemes

The elements of a Semitic word can be classified into four groups: a root, prefixes, infixes, and suffixes. A word, W, can be represented by the following:

$$W = P_1 + P_2 + \ldots + P_m + (R \mid I) + S_1 + S_2 + \ldots + S_n$$

where: P prefixes
R root
I infixes
S suffixes

Notes. (1)The operation | indicates the non-concatenative, non-linear merge of R and I.
(2) The values of m and n are language specific.

The unit (R | I) consists of two morphemes arranged in a non-concatenative, non-linear manner. R is the root morpheme which governs the basic notion of the word, excluding any inflectional information, e.g. «Syr. √ *ktb*» gives the notion 'write'. I is the inflectional morpheme which presents (partial) inflectional information, e.g. Syr. _ _*a*_ 'perf. act. m. 3 s.'. The unit (R | I), therefore, is the composition of R and I, e.g. «Syr. *ktab*» 'write + perf. act. s. 3 m.' or 'he wrote'. I shall call the unit (R | I) the *fundamental morphemes* (FM), 'fundamental' because they provide the two most crucial morphemes which govern the notion and the (partial) inflection of a word.

* This work is based on a project which was supervised by Dr Steve Pulman, Computer Lab., Cambridge.

Some inflectional information consists of prefixes, infixes and/or suffixes. For example, «Syr. *ketbat*» 'she wrote', whose *R* is also «Syr. *√ ktb*», contains the inflectional information *_e_ _at* 'act. perf. s. 3 f.', where *at* is viewed here as a suffix, given in the lexicon as *verbal inflectional marker* (VIM). Extracting the VIM *at* from *_e_ _at* yields that *I* consists only of *_e_ _*; this in turn means that the unit (*R* | *I*) consists of *ketb* which cannot stand on its own as a word.

Hence, I shall define two types of FMs. The first is free and can stand on its own as a word, thus *free FM* (FFM), e.g. «Syr. *ktab*» 'he wrote'. The second is bound and cannot stand on its own as a word, but requires an *inflectional marker* (IM) to compose a word, thus *bound FM* (BFM). The following formations of a word, *W*, are legal:

W = FFM
W = IM + BFM
W = BFM + IM
W = IM + BFM + IM

Further, I shall define two types of prefixes/suffixes: A *primary prefix / suffix* is essential to the formation of a word and is attached to a BFM, e.g. VIM. A *secondary prefix / suffix* is attached to a FFM, e.g. *w* in «Syr. *wketbat*» 'and she wrote'.

3. Linguistic Theoretical Background

3.1. Early Works on Semitic Morphology

The early studies on Semitic morphology were limited to the works of traditional grammarians whose main interest was to present Semitic to the student of Oriental languages. The grammatical literature of Western Orientalists on Arabic, Hebrew and Syriac is mainly based on medieval Semitic grammarians. The most detailed Western study in this regard was done by Brockelmann (1908-13). The most updated study on comparative Semitic morphology was done by Moscati *et al.* (1964).

The first linguistic account of Semitic morphology, from a modern linguistic point of view, was presented by Harris (1941) on the Hebrew of 600 B.C.; a second work was done by Chomsky (1951) based on transformational rule notation. McCarthy (1979) gave a more detailed study on the formal problems of Semitic phonology and morphology, and later (1981) proposed his prosodic theory of nonconcatenative morphology, which owes a great deal to Harris's work. Additional works by McCarthy and Prince (1990), and Farwaneh (1990) followed.

These studies form the linguistic background for any computational analysis of Semitic morphology and, therefore, are summarized below. (A discussion on Chomsky's work can be found in McCarthy (1981: 414-17).)

3.2. Harris's Analysis

Harris (1941) presented an analysis of the structure of the Hebrew current in Jerusalem and in official circles of Judah at about 600 B.C. His study was based on a reconstruction of Hebrew texts. Harris provided his analysis on different levels: phonemes, morphemes, words and phrases.

Morphemes, Harris stated, are of three classes: root morphemes, consisting of a sequence of consonants; pattern morphemes, consisting of vowels, or vowels and affixes; and a third class, consisting of successions of consonants and vowels. The latter is not of a great importance here.

The analysis of «Arab. *kuttib*», according to this study, produces two morphemes: the root morpheme *ktb* 'write' and the pattern morpheme *_u_i_* 'perfective passive' (: indicates the gemination of the middle radical). The word «Arab. *kuttib*» is the composition of the two morphemes.

3.3. *McCarthy's Prosodic Theory*

McCarthy (1981) did his analysis on the Arabic verb system. His study was done under the framework of autosegmental phonology (Goldsmith 1976).

McCarthy introduced three types of morphemes: consonantal root, consisting of the consonants of the root; vowel melody, consisting of a series of vowels; and templatic morpheme, consisting of a series of C's and V's called the CV-skeleton. Each of these three morpheme sits on a level of its own, and the three levels are coordinated by the principles of autosegmental phonology.

The analysis of «Arab. *kuttib*», according to this study, produces the following three morphemes: the root *ktb* 'write', the vowel melody /u-i/ 'perfective passive', and the templatic morpheme CVCCVC 'causative/factive'. The three levels are associated by the principles of autosegmental phonology as in Figure 1.

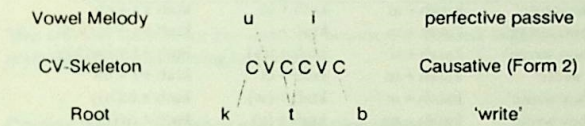


Figure 1.

When the universal rules of autosegmental phonology were not capable of representing the analysis of certain forms, McCarthy (1981) introduced language specific initial association rules which have special markings imposed on them.

Instead of imposing marking on the rules, Farwaneh (1990) argued that it is more appropriate to impose markings on the morphemes in the lexicon for two reasons: firstly, this explains regularities in some paradigms; secondly, it dictates that non-attested forms of a certain morphemes cannot be derived.

3.4. *Later Research*

McCarthy and Prince (1990) did further work on prosodic morphology. They argued that "templatic constraints on word structure should be characterized in prosodic terms". Instead of representing words using the three levels as shown in the example above, a vocabulary of the Prosodic Morphology Hypothesis, they argued, is more appropriate. Such prosodic terms are minimal word, foot, syllable and mora. Thus the word «Arab. *kuttib*» is analyzed as a sequence of two heavy syllables as illustrated in Figure 2.

3.5. *Limitations: The Case of Syriac*

The following table gives the paradigm of «P-S √ **ktb*» 'write' in the perfective active in Arabic and Syriac (letters between parenthesis are silent; ' + ' separates verbal inflectional markers from the verbal form). The last column gives the paradigms of Syriac followed by the singular-third-masculine object pronominal suffix.

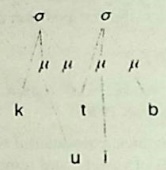


Figure 2.

	ENGLISH	ARABIC	SYRIAC	Syr. + OPS (s.3 m.)
s. 3 m.	'he wrote'	katab + a	ktab + **nil**	katb + eh
s. 3 f.	'she wrote'	katab + at	ketb + at	ktab + t + eh
s. 2 m.	'thou wrote'	katab + ta	ktab + t	ktab + t + oy(h)y
s. 2 f.	'thou wrote'	katab + ti	ktab + t(y)	ktab + t + i:w(h)y
s. 1 c.	'I wrote'	katab + tu	ketb + et	ktab + t + eh
p. 3 m.	'they wrote'	katab + u:	ktab + (w)	katb + u:(h)y
p. 3 f.	'they wrote'	katab + na	ktab + (y)	katb + oy(h)y
p. 2 m.	'you wrote'	katab + tum	ktab + tu:n	ktab + tu:n + oy(h)y
p. 2 f.	'you wrote'	katab + tunna	ktab + teyn	ktab + teyn + oy(h)y
p. 1 c.	'we wrote'	katab + na:	ktab + nan	ktab + n + oy(h)y

The analysis of the Arabic forms is clear: they consist of the root morpheme *ktb* 'write', the pattern *_a_a_* 'perfective active' (according to Harris) or the vowel melody /a/ 'perfective active' (according to McCarthy), and the templatic morpheme CVCVC (according to McCarthy). The analysis of Harris and McCarthy share one thing in common (apart from the root morpheme): both analysis extract the inflectional information 'perfective active' and assign it an independent morpheme (pattern and vowel melody, respectively).

The analysis of the Syriac forms (fourth column), however, causes some problems. The root morpheme *ktb* appears in all forms in the paradigm; however, it is not clear what constitutes the inflectional information 'perfective active' since s. 3 f. and s. 1 c. share a different inflection than the rest of the forms. Following Harris, there are two distinct patterns: *_ _a_* and *_e_ _*; following McCarthy, there are two vowel melodies: /a/ and /e/.

Before attempting to answer this question, a look at the syllabic structure is important. The VIMs of s. 3 f. and s. 1 c. start with a V, while the rest start with a C. Unfortunately, a complete survey of the syllabic structure in Syriac is not available, and I was unable to find a justification of the above phenomenon (Nöldeke (1904) offers no explanation). However, I searched for a CV sequence which is equivalent to the ill-formed *qtal + at* (CCVC + VC), and I found a nominal one for the word «Syr. *ktob + ok*» 'your book' which also satisfies (CCVC + VC)! No such verbal sequence exists.

As far as recognizing number, person and gender, scanning over the Arabic forms one can easily find that they are solely determined by VIMs (*a*, *at*, *ta*, etc.). This, however, does not hold for Syriac, where number, person and gender is determined by the inflection of the verb and the respective VIM.

The addition of suffixes causes additional problems. The last column gives the same Syriac verbs of the preceding column followed by the s. 3 m. OPS (when there are three elements separated by '+', the second is a VIP and the third is an OPS). It is worth noting that the form of the OPS is not the same in all occurrences. Leaving the form of the OPS for the moment, one still can see that the OPS caused the inflection of the verb to change in five places (s. 3 m., s. 3 f., s. 1 c., p. 3 m., p. 3 f.) in all occurrences of the paradigm. It is clear that the analysis of Harris and McCarthy do not take account of such problems.

For the analysis of Syriac, I shall adopt the following position: Inflectional semantics can have one-to-many relationship with inflectional patterns (vowel melodies). This means that 'perfective active' can be associated with all of the following vocalism: _ _*a*_ and _*e*_ _ when there is no OPS, and _*a*_ _ and _ _*a*_ when an OPS is required. In order to regulate these forms, unification-based features are used as in the following:

_ _*a*_	[OPS -]
_*e*_ _	[OPS -]
_*a*_ _	[OPS +]
_ _*a*_	[OPS +]

This solution suffices for the moment; however, a detailed linguistic study of such phenomena is necessary when considering large data.

4. Computational Theoretical Background

4.1. *Computational Morphology*

There are a number of approaches used in computational morphology. One of the well known systems is MITalk's DECOMP module (Allen et al. 1987) which is used in text-to-speech synthesis. Other systems were implemented for specific languages such as *keçi* for Turkish (Hankamer 1986).

The most influential approach to computational morphology, however, is the two-level approach which is not language specific and was successfully implemented on a variety of languages. It became the main-stream approach for current systems.

The following sections give a brief account of two-level morphology, followed by previous proposals for Semitic morphology. Finally, a formalism of two-level phonology is described.

4.2. *Two-Level Morphology*

The notion of two-level morphology was first introduced by Koskenniemi (1983). This approach defines two, and only two, levels of orthographic strings in recognition and synthesis: a lexical level and a surface level. The former is an orthographic representation of lexical entries; the latter is an orthographic representation of surface strings. A mapping scheme between the two levels was introduced and was implemented as FSTs.

Koskenniemi's two-level system was modified by various people. A number of proposals were made by Bear (1988), Trost (1990), Black et al. (1987) and Ruessink (1989). Due to limitations on the size of this work, these are not described here. A good account of Koskenniemi's system and the later developments can be found in Sproat (1992), and Pulman and Hepple (forthcoming).

Two-level morphology was implemented and tested on a number of languages. The only two-level implementation on a Semitic language, I am aware of, is Beesly (1990).

4.3. *Proposals for Semitic Computational Morphology*

This section gives a brief account of the few proposals presented for handling computational Semitic morphology. This account mainly deals with studies in the West. Work on Arabic morphology is taking place in the academic institutions of the Middle East, but unfortunately is not easy to get hold of; examples of such studies are El-Sadany (1989), Saliba and Al-Dannan (1989), and others.

Kay (1987) proposed a finite-state approach to nonconcatinative morphology using a FSM of four tapes, each containing one of the following: root, CV-skeleton, vowel melody and input. The first three constitute the three lexical morphemes of McCarthy (1981); the fourth is the surface string to be analyzed. Kay gives a full example of his proposal; an illustrative example is given in Sproat (1992). No implementation of Kay's proposal, according to my knowledge, has been done.

Kataja and Koskenniemi (1988) described a working system of Akkadian based on two-level morphology. The rules used are similar to two-level rules used in other languages. 'Interdigitation', i.e. nonconcatinativness, in this system puts more weight on the lexical structure. Two lexica are used: one for root morphemes, and the other for prefixes, infixes and suffixes. The lexical representation is defined as the intersection of the two lexica. Standard two-level rules describe the phonological and morphological processes, and are compiled into FST using the TWOL rule compiler (Karttunen et al. 1987). Unification-based features are used for morphosyntactic analysis.

Bird and Ellison (1992) proposed a system based on their description of one-level phonology. They defined a number of FSA for the following: form, root morpheme, and vocalism. A particular verb results by restricting the vocalism FSA to a second FSA which results from taking the product of the form FSA and the root morpheme FSA. This proposal assumes that the following suffixes are consonant-initial, and hence fails to describe a substantial number of the Syriac verbal forms as described above.

Pulman and Hepple (forthcoming) proposed a formalism for two-level phonology, and proposed using it for Arabic. This formalism was implemented here for Semitic. A description of their proposal is given below.

4.4. *Two-Level Phonology*

At the time when Koskenniemi (1983) was working on two-level morphology, Kay and Kaplan (1983) proposed a representation of phonology using FST. Later, the notion of two-level morphology gave birth to the notion of two-level phonology.

A formalism for two-level phonology was introduced by Pulman and Hepple (forthcoming). Phonology and morphology share things in common: "the two involve mapping between a surface representation of word forms onto a sequence of corresponding lexical morphemes". The main concern in this work is morphographemics rather than phonology. However, this formalism of 'phonology' is interesting for Semitic morphology for reasons which will be discussed later (section 5.4.2).

The formalism itself descends from earlier ones proposed by Black et al. (1987) and Ruessink (1989). There are two types of rules:

(1)	LeftSurfaceContext - LHS - RightSurfaceContext LeftLexContext - LHS - RightLexContext	= >
(2)	LeftSurfaceContext - LHS - RightSurfaceContext LeftLexContext - LHS - RightLexContext	< = >

Each of the above elements consists of sequences of zero or more segments. Variables over entire segments are allowed. The operator = > allows correspondences, while the operator < = > forces correspondences. The LHS and the RHS are the *active* parts of the rule. Examples of phonological processing is given in Pulman and Hepple (forthcoming).

Pulman and Hepple proposed applying this formalism to Arabic. For example, to represent «Arab. *kuttib*», the following rule can be applied:

* - C1 u C2 C2 i C3 - *	= >	0 - C1 C2 C3 - 0

where C1, C2 and C3 correspond to the three letters of a triliteral root morpheme, e.g. *ktb*, ('*' indicates no context).

5. The Computational Linguistic Framework

5.1. *Linguistic Framework*

The two available linguistic analysis of Semitic morphology are the previously discussed works of Harris (1941) and McCarthy (1981). The limitations of both analysis to describe some morphological Syriac phenomena was also discussed.

A computational application based solely on one of the two analysis does not seem feasible at the moment. The success of analysing a partial (or selective) class of forms may be doomed to failure when considering the whole paradigm of the language. Beside the limitations discussed above, the two analysis of Harris and McCarthy deal mainly with the singular third masculine verbal forms. An analysis of nominal forms is not provided by neither of them. In addition, their studies do not consider a wide range of Semitic languages.

This criticism by no means intends to put down the works of Harris and McCarthy, but aims to show that sticking strictly with just one of them may not be feasible. It must be stressed that any successful work on computational Semitic morphology would owe its success to the works of Harris and McCarthy. The relation of the formalism to these studies becomes apparent in the next section.

5.2. *Computational Framework*

Kay's (1987) approach is particularly attractive; its success remains to be seen when it is implemented on a large scale. Beesly (1990) proved that Koskenniemi's rules, in conjunction with 'detouring', are capable of implementing large Semitic systems. No details of 'detouring' are given (for intellectual-property reasons). Sproat (1992) gives his own interpretations of the method.

The formalism of Pulman and Hepple is of a particular interest for a number of reasons. Its most attractive feature, in terms of Semitic morphology, is that it is "not restricted to statements affecting pairs of surface and lexical segments, but can directly relate sequence, even sequences of different lengths." It is also very simple and clear formalism, a fact the grammar writer will be grateful for. However, the only way to handle the problems presented in Section 3.5 in this formalism is to duplicate rules with different features, which is not particularly efficient.

The problem of partial vocalization should be addressed. In Beesly's (1990) implementation, this problem is easily handled by allowing lexical vowels to surface with the same value or as nil:

Lexical Vowels Surfacing	Lexical Vowels Deleted
a:a	a:nil
u:u	u:nil
i:i	i:nil

This allows «Arab. *kutiba*» to surface as *ktb*, *kutb*, *kutib*, *ktib*, etc.

There are two ways by which this problem can be solved in the formalism of Pulman and Hepple. The first is adding rules to cover all possibilities. This, however, is not efficient and creates a huge grammar. Instead, variables can be used in the active LHS of the rule. For example, the rule representing «Arab. *kuttib*» can be given as follows:

* - C1 U C2 C2 I C3 - *	= >	0 - C1 C2 C3 - 0	where U in {nil, u} I in {nil, i}

This allows the LHS to match any of the following strings: *ktib*, *ktiib*, *kutib*, *kutiib*.

Pulman and Hepple point out that their formalism fails to distinguish between the three morphemes of McCarthy. They also question whether "the generalizations of McCarthy's account can be really maintained" when taking the whole paradigms of the language into consideration.

6. Conclusion

This paper gave a brief account of the linguistic and computational backgrounds of Semitic morphology. The most attractive formalism (for Arabic) is that of Pulman and Hepple. However, there is a need for a formalism which abstracts the inflectional information from the rules and allows them to reside in the lexicon in order to solve problems as those presented in section 3.5

7. References

- Allen, J., Hunnicutt, M., and Klatt, D. 1987. *From Text to Speech: The MITalk System*. Cambridge University Press.
- Bear, J. 1988. Morphology with Two level Rules and Negative Rule Features. In *COLING-88* (Association for Computational Linguistics).
- Beesley, K. 1989. Computer analysis of Arabic morphology: A two-level approach with detours. In *Proceedings of the Third Annual Symposium on Arabic Linguistics* (University of Utah).
- Beesley, K. 1990. Finite-State Description of Arabic Morphology. In *Second Cambridge Conference: Bilingual Computing in Arabic and English*, Cambridge.
- Bird, S. and Ellison, T. 1992. One Level Phonology: Autosegmental Representations and Rules as Finite-State Automata. In *Edinburgh Research Papers in Cognitive Science*, No. EUCCS/RP-51, University of Edinburgh.
- Black, A., Ritchie, G., Pulman, S., and Russel, G. 1987. Formalism for Morphographemic Description. In *ACL Proceedings, 3rd European Meeting* (Association for Computational Linguistics).
- Brockelmann, C. 1908-13. *Grundriß der vergleichenden Grammatik der semitischen Sprachen*, 2 vol., Berlin.

- Chomsky, N. 1951. *Morphophonemics of Modern Hebrew*, Master's thesis, University of Pennsylvania, Philadelphia.
- Farwaneh, S. 1990. Well-Formed Associations in Arabic: Rule or condition? In ed. M. Eid and J. McCarthy *Perspectives on Arabic Linguistics II*, John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Goldsmith, J. 1976. *Autosegmental Phonology*, Doctoral dissertation, MIT, Cambridge, Massachusetts.
- Hankamer, J. 1986. Finite state morphology and left to right phonology. In *Proceedings of the West Coast Conference on Formal Linguistics*, Volume 5 (Stanford Linguistic Association).
- Harris, Z. 1941. Linguistic Structure of Hebrew, *Journal of the American Oriental Society* 61.
- Kataja, L. and Koskenniemi, K. 1988. Finite-state description of Semitic morphology: A case study of ancient Akkadian. In *COLING-88* (Association for Computational Linguistics).
- Karttunen, L., Koskenniemi, K., and Kaplan, R. 1987. A Compiler for Two-level Phonological Rules. In *Tools for Morphological Analysis*, Center for the STudy of Language and Information, Report No. CSLI-87-108.
- Kay, M. 1987. Nonconcatenative finite-state morphology. In *ACL Proceedings, 3rd European Meeting* (Association for Computational Linguistics).
- Kay, M. and Kaplan, R. 1983. Word Recognition. Manuscript, Xerox Palo Alto Research Center.
- Kiraz, G. 1992. *Semitic Languages and Two-Level Morphology*. M.Phil. thesis, University of Cambridge.
- Koskenniemi, K. 1983. Two-Level Morphology: *A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki.
- McCarthy, J. 1979. *Formal Problems in Semitic Phonology and Morphology*, Doctoral dissertation, MIT, Cambridge, Massachusetts.
- McCarthy, J. 1981. A Prosodic Theory of Nonconcatenative Morphology, *Linguistic Inquiry* 12, no. 3.
- McCarthy, J. 1990. Prosodic Morphology and Templatic Morphology. In ed. M. Eid and J. McCarthy *Perspectives on Arabic Linguistics II*, John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Moscati, S., Spitaler, A., Ullendorff, E., von Soden, W. (1964). *An Introduction to the Comparative Grammar of the Semitic Languages, Phonology and Morphology*, Porta Linguarum Orientalium, Neue Serie VI. (Otto Harrassowitz, Wiesbaden.)
- Nöldeke, T. 1904. *Compendious Syriac Grammar*, translated by James Crichton, London.
- Pulman, S. and Hepple, M. Two Level Morphology. Forthcoming in *Computer Speech and Language*.
- Ruessink, H. 1989. Two Level Formalism. *Utrecht Working Papers in NLP*, No 5.
- El-Sadany, T. 1989. *Arabic Morphological Analyzer on a personal computer*. M.Sc. thesis, University of Al-Azhar, Egypt.
- Saliba, B. and Al-Dannan, A. 1989. Automatic Morphological Analysis of Arabic: A Study of Content Word Analysis. In *First Kuwait Computer Conference*.
- Sproat, R. 1992. *Morphology and Computation*. MIT Press, Cambridge, Massachusetts.
- Trost, H. 1990. The Application of Two level Morphology to non-concatenative German morphology. In *COLING-90* (Association for Computational Linguistics).

George A. Kiraz
Peterhouse, Cambridge CB2 1RD

G. Kiraz holds a B.Sc. in Engineering from California State University, an M.ST. in Syriac Studies from the University of Oxford, and an M.Phil. in Computer Speech and Language Processing from the University of Cambridge.