Finite-state Description of Semitic Morphology: A Case Study of Ancient Akkadian

Laura KATAJA
University of Helsinki
Department of Asian
and African Studies
Hallituskatu 11
SF-00100 Helsinki
Finland

Kimmo KOSKENNIEMI
University of Helsinki
Research Unit for
Computational Linguistics
Hallituskatu 11
SF-00100 Helsinki
Finland

Abstract: This paper discusses the problems of description and computational implementation of phonology and morphology in Semitic languages, using Ancient Akkadian as an example. Phonological and morphophonological variations are described using standard finite-state two-level morphological rules. Interdigitation, prefixation and suffixation are described by using an intersection of two lexicons which effectively defines lexical representations of words.

1. Introduction

Word-formation in Semitic languges poses several challenges to computational morphology. One obvious difficulty is its nonconcatenative nature ie. the fact that inflection is not just adding prefixes and suffixes, but also includes *interdigitation* where the phonological sequence symbolizing a verbal root is interrupted by individual and short sequences of phonemes denoting various derivational and inflectional stems. In addition to this, there are numerous phonological and morphophonological processes of a more conventional character.

Two-level phonology assumes a framework for word-formation where there is an underlying lexical representation of the word-form and a surface representation which are related to each other with two-level rules [Koskenniemi 1983]. These rules compare the representations *directly* and they operate in *parallel*. The lexicon component defines what lexical representations are permissible and how they correspond to sequences of morphemes, see figure 1.

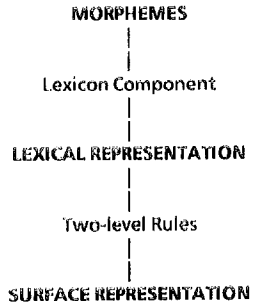

Fig. 1

This paper describes a fairly comprehensive two-level rule system for phonological and morphophonological alternations in Akkadian word inflection and regular verbal derivation. The rule component proves to be similar to two-level rule systems for other languages.

Interdigitation entails more requirements for the lexicon which defines feasible lexical representations and relates them to underlying morphemes. The task for the lexicon component is more or less universal, even if some languages can do with simpler lexicons while others require more sophisticated structures.

This paper discusses a solution which involves using two separate lexicons, one for word roots, and the other for prefixes, flexional elements and suffixes. Entries for roots leave flexional elements unspecified and vice versa. The intersection of these two lexicons effectively defines lexical representations of word-forms.

2. Morphotactic structure of word-forms.

Akkadian verbs have the following overall pattern:

[pers.] [root & flection] [gender & numb.] [opt. subjunctive etc.] [opt. obj.]

An example of a full fledged verbal form would be *'(that) they caught him'*:

lexical representation:	i	X	t	a	B	A	T	-	u	\	-	n	i	- sh u
surface representation:	i	x	x	a	b		t		u	u		n	i	sh u

A dash '-' denotes morpheme boundary, and backslash '\' a morphophoneme for vowel lengthening. The above word-form is divided into its parts according to the pattern as follows:

person	i
root	X ... B ... A T
flection	... t a ... ...
gender & number	u \
subjunctive	n i
object	sh u

Capital letters are used in order to distinguish radical consonants and vowels from segments in other morphs. Thus, the root & flection part is XtaBAT where capital letters are components of the root, with lower case letters representing flectional elements.

Nouns, in turn, have an overall structure :

[stem] [case & number] [opt. possessive]

An example of a maximal nominal word-form is *'their kings'*:

lexical representation:	Sh	a	R	\	-	a	\	n	i	-	sh	u	n	u
surface representation:	sh	a	r	r		a	a	n	i		sh	u	n	u

This can be readily decomposed into its parts as follows:

stem	Sh a R \
case & number	a \ n i
possessive	sh u n u

3. Overall structure of morphs

Verbal roots have an overall pattern of three radical consonants and one vowel C ... C ... V C where flectional elements may occur in the two intervening slots marked with "..."

Flectional elements have a pattern consisting of two parts to fill the corresponding two gaps in the verbal root. The overall pattern is roughly ...(((C)C)V)...(V or \)...

There is at most one *verbal prefix* and it indicates person (and partly modus). Its overalll pattern is (C)V.

There are at most three *verbal suffixes* attached to the stem. The first suffix indicates gender and number (and partly person). They have the form V\ or they are empty. The second suffix indicates either the subjunctive (u, empty, or ni) or the ventive (am or nim). The third suffix denotes the object or the dative case and conforms to a pattern C V(C V(\ C V))

Nominal stems are given as derived complete stems containing three radical consonants which can be identified, but no attempt has been made to generate them from plain radical consonants and flectional elements because stems are idiosyncratic and better described as lexicalized whole units.

Nominal suffixes indicate gender, number and case. Gender is part of the stem for nouns whereas adjectives have an explicit feminine suffix (a)t (the masculine has no marking). Number and case are represented by portmanteau morphs. After these endings there may be a possessive ending according to one of two patterns: V \ or C V (C V).

3. Phonological Description

Akkadian, like many other Semitic languages, has a considerable number of phonological and morphophonological processes. This paper describes a fairly complete and tested system of some 30 rules written in two-level formalism and compiled with the TWOL rule compiler [Karttunen, Koskenniemi and Kaplan, 1987]. A number of examples is given below accompanied by rules that correspond to the processes. In each example the lexical representation is given (in bold face) above the surface representation (in normal face).

There are several assimilations word internally and at morpheme boundaries, eg. an N in the root is assimilated to the immediately following consonant, eg. *'he cut (past tense)'*:

```
i N K I S
i k k i s
```

which corresponds to the rule:

```
"assimilation of N"
    N:F  <=>  __ :F ; where F in Consonants ;
```

Futhermore, *'he said'*:

```
i Z t a K A R
i z z a k a r

"assimilation of dentals"
    t:F  <=>  :F __ ; where F in Dentals ;
```

and *'he trusted him (something)'*:

```
i P Q I D - sh u
i p q i s   s  u

"suffix assimilation of t"
     t:s  <=>  __ -: sh: ;

"suffix assimilation of sh"
     sh:s  <=>  :s -: __ ;
```

Some alternations caused by laryngeals:

'lord'

```
B a \ E l u
b e e   l u

"umlaut"
    a:e   =>  E: :* __ ;
                 __ :* E: ;
```

'he enters'

```
i E a R \ U B
i     r r u b

"elision of a"
    a:0  <=>  :Vowel Laryngeal: __ :Consonant \:
;
```

Examples of deletion of short vowels:

'good'

```
D a M i Q u
d a m   q u
```

Examples of vowel contractions.

'they said to me'

```
i Q B I J u \ - n i m - n i
i q b     u u   n i n   n i

"Vowel contraction"
    Vowel:0  <=>  __ (La:) :Vo (La:) :Vo ;
```

'(she is) clean'

```
Z a K U J - a t
z a k a     a t
```

Examples of morphological alternation of root vowels:

'he decides'

```
i P a R \ O S
i p a r r a s
```

'decide!'

```
  P v R O S
  p u r u s
```

Some analogical forms:

'he enters'

```
i E a R \ J B
i     r r u b
```

'they (fem.) donate'

```
i Q a I \  A1 Sh a1 \
i q   i sh    sh a  a
```

4. Structure of the lexicon

Lexicons are often understood as lists of entries or as some kind of tree structures having branches with letters as their labels (tries). A tree is, of course, a special case of a finite-state transition diagram or a finite-state automaton. Specifically, trees have no loops or cycles. The obvious generalization of lexicons would, then, be to use transition diagrams instead of trees. An entry for a verbal stem *'decide'* as a regular expression could be:

$$\Sigma_2^* \; p \; \Sigma_2^* \; q \; \Sigma_2^* \; o \; s \; \Sigma_2^*$$

where Σ_2 denotes the alphabet for prefixes, flectional elements and suffixes. Correspondingly, an entry for a present tense basic stem (G stem) could be:

$$\Sigma_1 \; a \; \Sigma_1 \; \backslash \; \Sigma_1 \; \Sigma_1$$

where Σ_1 denotes the alphabet for radical consonants and vowels. Intersections of such root entries and flectional elements give exactly the lexical representations of verbal stems. (The number of different entries needed for flectional parts is in the order of 70.)

The inflectional part of the lexicon could be expressed as a concatenation of the prefix, flexion and the suffix sub-lexicons. The intersection of this and the root lexicon contains all feasible lexical representations (which was the task of the lexicon component). This intersection need not be carried out in advance because the process of recognition can perform simultaneous searches in these two components and simulate the intersection. The result of an actual intersection would be inconvenient because of its size (roughly, the product of the sizes of its components). (There is no operational implementation of this part of the system yet, although facilities to build it are available.)

5. Combinations of Morphemes

The structure of lexicon that was sketched above greatly overgenerates, because many combinations of prefixes, flectional elements and suffixes are not valid. Restrictions are needed for the cooccurrence of these morphemes. One obvious way to cope with such combinatorics is to use unification-based features as in D-PATR [Karttunen 1986]. Unification features have the additional benefit of also providing effective morphosyntactic features for word-forms. It seems that the ability of using negation and disjunction in unification would simplify the description. In the following we assume these to be avaliable.

Effective restrictions for prefixes could be eg.:

u	(notPers2 singular notFemin) or (person3 plural notComm)
i	(person3 singular masculine) or (person3 plural notComm)

where Comm refers to a gender which is used in some forms to cover both feminine and masculine (feminine, masculine and Comm are mutually exclusive).

Descriptions for suffixes could be eg.:

a a	notPers1 plural notMasc
i i	person2 singular feminine
u u	person3 plural masculine
null morph	notPers2SgFem or person1 plural comm

The templates can be defined in a straight forward manner to result in combinations eg.:

u ... a a	person3 plural feminine
u ... u u	person3 plural masculine
u ...	person1 singular masculine, or person3 singular masculine

The combinatorics of Akkadian prefixes and suffixes seems to be fairly complicated, but a feature calculus seems to be sufficient for handling it so that it lets only valid combinations through and gives correct morpho-syntactic features to word-forms. (This part of the work is still in progress.)

References

Karttunen, Lauri (1986) "D-PATR: A Development Environment for Unification-Based Grammars", *Proceedings of COLING'86*, Bonn.

Karttunen, L., Koskenniemi, K., 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, Martin, [an unpublished paper on Finite-State Aproach to Arabian Morphology at a Symposium on Finite-State Phonology at CSLI in July 1985.]

Koskenniemi, Kimmo (1983) *Two-Level Morphology: A General Computational Method for Word-Form Recognition and Production.* University of Helsinki, Department of General Linguistics, Publicatiuons, No. 11.

Koskenniemi, Kimmo (1986) "Compilation of Automata from Morphological Two-level rules", *Papers from the Fifth Scandinavian Conference of Computational Linguistics.* Helsinki, December 11-12, 1985. Department of General Linguistics, Publications, No. 15.