Writing Kimmo Lexicons

1.0 First, two general tips on rule writing and lexicon writing.

Besides tracing a rule, there are two other PCKIMMO commands
that you can use.

1. The first is SET RULE <rule number> {ON|OFF} this lets you turn individual rules on or off. This is very helpful if you find that your recognizer returns NO results - ie, gives the output ****NONE***** when it should be giving some parse. Then you can turn off rules one by one, and you can isolate an offending rule this way. (Usually, a rule that has the system winding up in a nonfinal state when it should not be..)

To use this, your rules must beging with numbered double quoted comments lines as in,

"Rule 3 EPTHENSIS.... mumble....." 5 6

The only thing that matters here in quotes is the "Rule 3" business.

If you want to see your list of rules, do SHOW RULES

2. The second debugging aid is the SHOW RULE <rule number> command This is very helpful in dealing with one of the trickier parts of writing rules: how the different subsets and feasible pairs (lexical/surface characters) interact, even within one rule. SHOW RULE will tell you what characters are ACTUALLY being processed by the automata, which can differ from what you wrote down in the automaton spec.

Let me illustrate with an example.

Consider the following INCORRECT Epenthesis rule. Here S represents the subset s, x, z. Remember this is the rule that is supposed to pair fox+s with foxes. That is, IF the the pair 0:e appears, then it must have a left context of S (where S is x,z, s) followed by +:0 and the right context s#. (note that the right context doesn't care what the underlying lexical form is, so we don't write it down.) That is a DECLARATIVE CONSTRAINT that says this pair is OK - note that the left and right context of 0:e is indeed one of the members of S:S (in this case, X:x). followed by +:0. And the right context is indeed simply s # (on the surface -- we don't have to mention the hash mark # boundary symbol in the lexical or underlying string, really - it is assumed to be the same as the surface string.) So we are really lining up the following pair of characters, where I have written the surface characters in lower case. This is, in fact, how you can develop your own automata. First try pairing up lexical and surface strings, for the Spanish examples.

```
     F  O   X   +    0   S    #            (lexical, or underlying)
     f  o    x   0    e   s    #           (surface)

RULE   "3  Epenthesis.  0:e ==> S  +:0____s#"  5 6
      s   S   +  0  #  @
      s   S   +  e  #  @
1:  1   2   1  0  1  1
2:  1   2   3  0  1  1
```

```
3:  1   2   1  4  1  1
4.  5   0   0  0  0  0
5.  0   0   0  0  1  0
```

The 5, 6 at the end of the RULE statement gives us the number rows and columns in the state table. Recall that 0 here means a reject state. The states are listed on the leftmost column. The transition arc labels are the top row of (lexical, surface) pairs - feasible pairs. The inner cells say what the next states are - if we are in state 1, and see an s/s, then we go to state 1.

Given the lexical form fox+s, this table correctly produces foxes, but given kiss+s, it fails to produce the form kisses. Doing the SHOW RULE command will give us the following information, to tell us why it fails:

```
>show rule 3

3  on Epenthesis   Epenthesis.  0:e ==> S  +:0____s#"

s:s   (  s:s  )
S:S  (  x:x   z:z   )
+:0  (  +:0 )
0:e   (  0:e  )
#:#   (  #:#  )
@:@  (  b:b  d:d  F:F  g:g  j:j  k:k  l:l  m:m  n:n  p:p  q:q  r:r  t:t
    v:v  w:w  y:y  a:a  e:e  i:i  o:o  u:u  ':'  -:-  -:0  ':0  )
```

From this display, it is obvious that the column header S:S does NOT contain the pair s:s as might be expected. This is because the column headers s:s and S:S OVERLAP with respect to the pair s:s --- this pair matches BOTH. The pair s:s is assigned to the s:s column because that one is MORE SPECIFIC than the S:S column header. That is, only ONE feasible pair matches the s:s header, while three pairs match the S:S header.

Thus the input form kiss+s FAILS the fule because the final s in the root "kiss" is matched to the s:s column, leaving the table in state 1. -- where we really want it to be in state 2 (the left context mentions S:S, that is what state 2 is doing for us...) The table must be revised so that for the first three states the s:s column has the same state transitions as the S:S column.

Finally, the command SHOW LEXICON <lexicon name> - this will help if you mistype a lexicon name, etc.

2.0 Writing KIMMO lexicons.

(my apologies if you received this more than once - something went amiss with the AI mailer. Also, I've posted this material on the web site).

Here we explain how two-level rules work, how they can be implemented as finite-state machines, and all the types of rule constraints can be translated into finite-state tables. We then summarize the rule semantics. This is followed by a detailed discussion of rule conflicts; specficity and conflicts amongst SUBSETS; and finally, and explanation of the rule file format and the rules

in the pc-kimmo file english.rul.

It's a lot to read through... but I hope, complete,
and will guide you through Spanish.

1. How two-level rules work.

Consider Rule 2 (R2) below.

R2   t:c ==> ____i

The operator ==> means that lexical t is realized as a surface c only
(but not always) in the environment preceding i:i.

The correspondence t:c declared in R2 is a special correspondence.
All two-level descriptions must also contain a set of *default*
correspondences, such as t:t, i:i, etc. (This is the so-called
"BOGUS RULE" - it isn't really bogus, it is a default.)
The sum of the special and default correspondences are the total set
of valid correspondences or feasible pairs that can be used in the
description.

If a two-level description containing R2 (and all default
correspondences) is applied to the lexical (underlying) form "tati"
(without the quote marks) PCKIMMO proceeds as follow to produce the
corresponding surface form(s). (NOTE this is why you can use
GENERATE without a dictionary and JUST the .rul file)

Beginning with the first character of the input form, it looks to see
if there is a correspondence declared for it. Due to R2, it will find
that lexical t can correspond to surface c, so it will begin by
positing that correspondence.

```
Lexical:   t   a     t   i
           |   |     |   |
Rule:     R2
           |
Surface:   c
```

At this point the generated has entered R2. For the posited t:c
correspondence to succeed, the generator MUST find an i:i
correspondence next - that is what R2 says. When the generator moves
on to the second character of the input word, it finds that it is a
lexical a, and thus R2 FAILS, so the generator must back up, undo
what it has done so far, and try to find a different path. Backing up
to the first character t, it now tries the DEFAULT correspondence t:t
(which is guaranteed to succeed, since it has NO conditions):

```
Lexical:   t   a     t   i
           |   |     |   |
Rule:     R2
           |
Surface:   t
```

The generator now moves on to the second character. No correspondence
for lexical a has been declared other than the default, so the
generator posits a surface a:

```
Lexical:   t   a     t   i
           |   |     |   |
```

```
Rule:     R2   |
           |   |
Surface:   t   a
```

Moving on to the third character, the generator again finds a lexical t, so it posits a surface c and enters R2 again:

```
Lexical:   t   a     t   i
           |   |     |   |
Rule:     R2   |    R2
           |   |     |
Surface:   t   a     c
```

Now the generator looks at the fourth character, a lexical i. This SATISFIES the environment of R2, so it keeps the i (NOTE that the constraint refers only to a surface i, and says nothing about the lexical, underlying character):

```
R2   t:c ==> ____i
```

Since the context of R2 requires an i, the generator must also posit a surface i, so it does, and exits R2. NOTE that by the time R2 is finished, TWO characters will have been posited.

```
Lexical:   t   a     t   i
           |   |     |   |
Rule:     R2   |    R2   |
           |   |     |   |
Surface:   t   a     c   i
```

Since there are no more characters in the lexical form, the generator outputs the surface form "taci". However, the generator is not yet done. It will continue backtracking, trying to find alternative realizations of the lexical form. First, it will undo the i:i correspondence of the last character of the input word, then it will consider the third character, lexical t. Having already tried the correspondence t:c, it will try the default correspondence t:t:

```
Lexical:   t   a     t   i
           |   |     |
Rule:     R2   |     |
           |   |     |
Surface:   t   a     t   i
```

Now the generator will try the final correspondence and succeed, since R2 does NOT prohibit t:t before an i (rather, it prohibits t:c in any environment EXCEPT BEFORE i). It will then output "tati". The reader may confirm that no other outputs will be generated.

2. The ==> rule as a finite-state machine.

A key insight of PCKIMMO is that if phonological rules are written as two-level rules, they can be implemented as FST's running in parallel. In the next 4 sections we briefly show how each of the four rule types (==>, <==, <==>, and \<==) translates to an FST. We then go on to describe conflicts in SUBSETS, and RULES.

2.1 A ==> rule.
Consider rule R2 again.

A possible paraphrase is, If ever the correspondence t:c occurs, it must be followed by i:i. In other words, if anything OTHER THAN t:c occurs, this rule ignores it. This must be incorporated into our two-level FST, call this T2 (for table 2)

	t	i	@
	c	i	@
1:	2	1	1
2.	0	1	0

The @:@ arc means ANY OTHER symbol than t, i, or c, i.
State 2 is a kind of 'default'state that ignores everyting except the substring crucial to the rule. It is also the only final, accepting state.

Importantly, the state table is constructed such that the entire set of feasible pairs in the rule description is partition among the column headers WITH NO OVERLAP (this is the source of MANY bugs in Kimmo rule systems). T2 specifies the special correspondence t:c and the environment in which it is allowed. (the machine goes to state 2 to anticipate that an i:i comes next - if it does, success, and goes to state 1; if not, it goes to state 0, the rejecting state.)

The column header @:@ in T2 matches ALL the feasible pairs that are defined by ALL THE OTHER FSTs of the system - thus saying that R2 'takes a pass' and doesn't care about any other feasible pairs. So, with respect to T2, @:@ does not stand for all feasible pairs, rather, all feasible pairs except i:c and i:i.

The default correspondences of the system must be declared in a trivial FST like T3: (also see below where we cover the .rul file format). If we assume p, t, k, a, i, u in our alphabet, then we need:

	p	t	k	a	i	u	@
	p	t	k	a	i	u	@
1:	1	1	1	1	1	1	1

(Table T3)

Even this table of correspondences must include @:@ as a column. Otherwise, it would fail when it encountered a special correspondence such as t:c, because all the rule in a two-level description apply in parallel, and for each character in an input string ALL the rules must succeed, even if vacuously. Now, given the lexical form tatik, T2 and T3 together will generate the surface forms tatik and tacik.

IMPORTANT. To understand how to represent two-level rules as state tables, we must understand what the rules really mean. It is a common tendency to think of them positively, that is, as statement of where the correspondence succeeds. IN FACT STATE TABLES ARE FAILURE DRIVEN, THEY SPECIFY WHERE THE CORRESPONDENCES MUST FAIL.

This point is perhaps THE biggest source of difficulty in building the FSTs.

In our case above, it is natural to think of R2 as saying that t:c succeeds when it occurs preceding i:i. But T2 actually works because

it FAILS when ANYTHING BUT i:i follows t:c.

2.2 A <== rule.

Now consider R4.

R4 t:c <== ____i

This rule says that lexical t is always realized as surface c when it occurs before i:i, but NOT ONLY BEFORE i:i. Thus, the lexical form tati will successfully match the surface form taci, but not tati. Note, however, it would also match "caci" since it does not disallow t:c in any environment. Rather, its function is to disallow t:t in the environment following i:i.
Remember that state tables are failure-driven, so the strategy of writing the state table for R3 is to force it to fail if it recognizes the sequence t:t i:i. So the state table for R4, viz., T4, looks like this:

T4

	t t	i	@
	c t	i	@
1:	1 2	1	1
2:	0 2	0	1

In state 1, any occurrences of the pairs t:c, i:i, or any other feasible pairs are allowed without leaving state 1. It is only the correspondence t:t that forces a transition to state 2, where all feasible pairs succeed except i:i. Note that state 2 must be a final state - this allows all the correspondences to succeed and return to state 1. Also note that in state 2 the cell under the t:t column contains a 2. This is necessary to allow for the possibility of a tt sequence in the input. For example, tatti will surface as the form tatci. This phenomenon is called "backlooping" - more on this below.

Actually T4 is potentially over-specified. It is not really the pair t:t that is disallowed before i, but rather the pair t:not-c (lexical t and surface anything but c) Given that the more specific correspondence t:c is already in the table, the more general correspondence t:@ will take care of all the rest of the characters, including t:t. (I'll leave the details of this to you..)

In summary, the rule type L:S <==E positively says that L is ALWAYS realized as S in the environment E. Thus, it is a kind of OBLIGATORY rule. Negatively, it says that L is realized as any character but S is not allowed in E. The state table must be written so that it forces all correspondences of L with anything BUT S to fail.

2.3 A <==> rule.

R5 t:c <==> ____i

The state table for a <==> rule is simply the combination of the tables for ==> and <==. You build it by anding the two fst's together. So here, t:c MUST occur before i, and NOWHERE ELSE.

We next turn to the problem of what can happen when you have more than one rule - rule conflicts, the use of SUBSETS, and overlapping character descriptions.

[this part on rule conflicts, subsets, and an illustration via
the english.rul file will be sent next]