Slide 1

Automated Transliteration

Serge ROSMORDUC

Équipe EC.ART, laboratoire LRIA, Université Paris 8

Slide 2

Plan

1. Introduction
2. Interest of Transliteration
3. Principles of automated transliteration
4. An improvement: transducers
5. Architecture of the system
6. Some examples
7. Comments on the system
8. Future tracks

Slide 3

Introduction

1. Context
2. Interest of Transliteration
3. Points to take into account

Slide 4

Context

The tksesh Software

Slide 5

Interest of Transliteration

- User convenience
- Eases searching
- Word analysis and comparison
- Intellectual challenge

Slide 6

Points to take into account

- Sign values
- Sign combinations
- Word composition
- Word length
- Grammatical words
- Group-writing and “ligatures”
- Signs with peculiar behaviour

Slide 7

Principles of automated transliteration

Rewriting rules

Word to analyse : 𓌸𓃀𓀁

Made of signs :

𓌸 P(A,b) or P(m,r)

𓃀 P(b)

𓀁 DET(mouthAction)

Slide 8

Rules :

```
a) P($X, $Y), P($Y) => L($X), L($Y) / 150
b) P($X) => L($X) / 380
c) P($X, $Y) => L($X), L($Y) /  400
d) DET($X) => DET($X)
```

choices :

1. a and d : *ꜣb*
2. b, c and d : either *ꜣbb* or *mrb*, depending on sign values used

Costs : first hypothesis : 150, second 780. First wins.

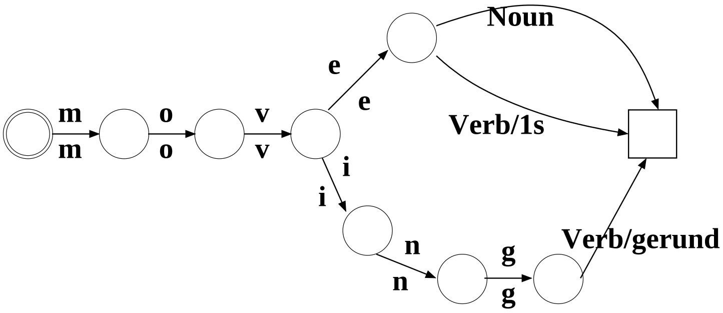
Slide 9

Comments on the method

- implemented. Works reasonably well.
- unable to cope with some phenomena :
  - word cutting, word length
  - composed signs, groups of signs
  - doesn't work well with so-called phonetic determinatives.

Slide 10

An improvement: transducers

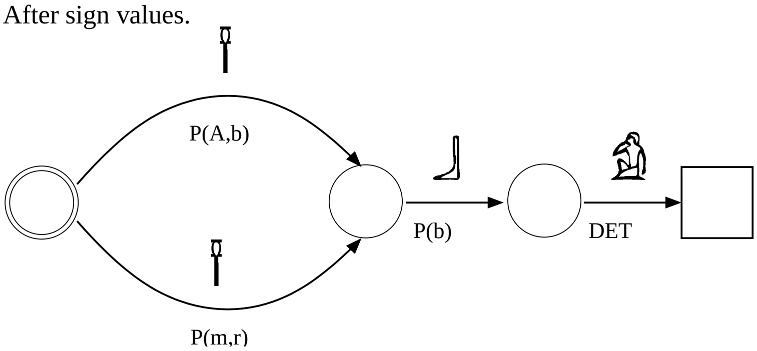


- Efficient.
- Can represent many hypothesis in a compact way.
- Can be composed.

Slide 11

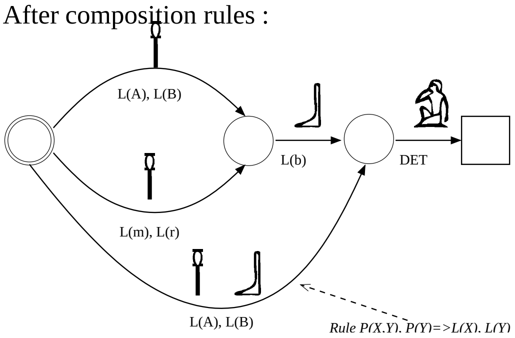
Simple example : 𓍋𓃀𓀁

After sign values.



Slide 12

After composition rules :



Slide 13

Architecture of the system

1. Entry ;
2. normalization ;
3. Word limits markers ;
4. Sign values ;
5. First combinatory rules ;
6. Second combinatory rules ;
7. Word length rules.

Slide 14

normalization

(from MdC to Gardiner codes). Variant codes are also normalized.

```
Y1 => Y1
Y1v => Y1
mDAt => Y1
Y2 => Y1
```

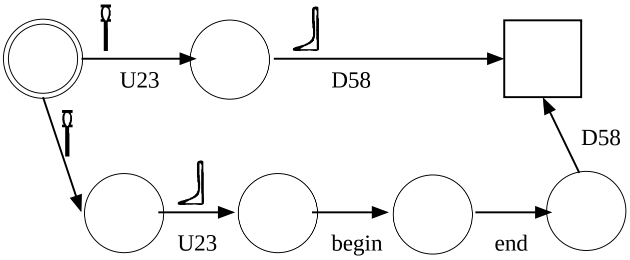
Slide 15

Word limits markers

Word beginning and endings are explicitely marked.

Inserts the possibility of a word break between each pair of signs.

```
$X => $X
$X, $Y => $X, end, begin, $Y
```



Slide 16

Sign values

Propose values for signs :

```
U23 => P(A,b) / 10
U23 => P(m,r) / 10
```

Also place for groups of signs, group writing, etc.

```
F9, F9 => P(p,H,t,y) / 10
# nsw
M23, X1, N35 => IP(n,s,w) / -10000
G20, D36 => P(m) / 10
I3, I3 => IP(i,t,y) / 10
M17, M17 => P(y) / 10
```

Slide 17

Values are :

- P(X,Y) : phonetic sign
- IP(X,Y) : phonetic determinatives
- ID(X,Y,Z) : ideogram
- DET(X) : determinative
- NUM(X) : numeric
- W(X) : monogram
- END(X) : Z1 or Z3

Slide 18

First combinatory rules

Combine signs at a “local” level. Produces likely word endings, and phonetic strings (L(X)).

```
P($X, $Y, $Z), P($X, $Y) => L($X),
                               L($Y), L($Z) / 60
DET($X), fin => DET($X), fin, R(46) / 0
DET($X), END($E), fin => DET($X), fin, R(145) / 0
```

Slide 19

Second combinatory rules

Combines the phonetic strings from the first set to remaining signs, in particular phonetic determinatives.

```
L($X), L($Y), L($Z), IP($X,$Y,$Z) =>
              L($X), L($Y), L($Z) / -2000
```

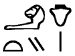
Slide 20

Word length rules

word length	cost
0	100000
1	200
2	100
3	0
4	210
5	1600
6+	1600 + (n-5)*800

Slide 21

Some examples

- vs. 
- the shipwrecked sailor.

Slide 22

Comments on the system

- rather good results ;
- flexible : for different kind of texts, change the rules ;
- rules cost are difficult to assess and to change ;
- Problems with grammatical words ;
- Solutions lie in closer study.

Slide 23

Future tracks

Group system : a word is made of a prefix, a core, and a suffix.

No more word length rules: implied by the finer control.

Allows to explain that “w”, “y”, “t”, are likely in word endings.

Will create a very structured representation of the words. Interesting for searches.

Detailled analysis needed for progress.