Translittération automatique des hiéroglyphes égyptien

S. Rosmorduc
rosmord@iut.univ-paris8.fr
équipe « langue et littérature de l'Égypte ancienne »

Introduction

- Les logiciels Tksesh/JSesh
  - but : environnement de travail pour le philologue
  - éditeur et base de textes
  - éditeur de signes
  - lexique
  - multi utilisateurs
  - en cours de réécriture en java

Principes de l'écriture hiéroglyphique

signes consonnantiques

unilitères  m

bilitères  $3b$

trilitères $ḥpr$

idéogrammes

déterminatifs ; ;

Exemples

 *sb3*, enseigner

 *sb3*, étoile

 *idem*

 *sb3*, porte

 idem.

Caractéristiques importantes pour l'informatisation

- signes à valeurs multiples
- système redondant
- orthographes variables
  - diachroniquement
  - synchroniquement
  - selon support

Mise en place de règles de réécriture

Mot à analyser : 

Composé des signes :

 P(*ȝ*,*b*) ou P(*m*,*r*)

 P(*b*)

 DET(actionBouche)

règles de réécriture (suite)

Règles:

```
a) P($X, $Y), P($Y) => L($X), L($Y) / 100
b) P($X) => L($X) / 100
c) P($X, $Y) => L($X), L($Y) /  100
d) DET($X) => DET($X)
```

Choix possibles

1. a et d : *3b*
2. b, c, et d : *3bb* ou *mrb,* selon la valeur du signe.

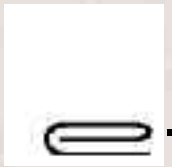
Coût : première hypothèse 100, seconde 200 La première l'emporte.

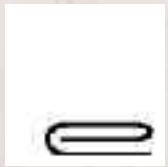
Le problème de l'explosion combinatoire

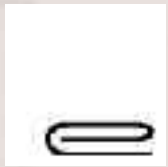
 *wḏb, tourner, plier*

 -> P(*b*)

-> P(*w,ḏ)*

 -> P(p,X,r)

 -> IP(w,D,b)

 -> DET(entourer)

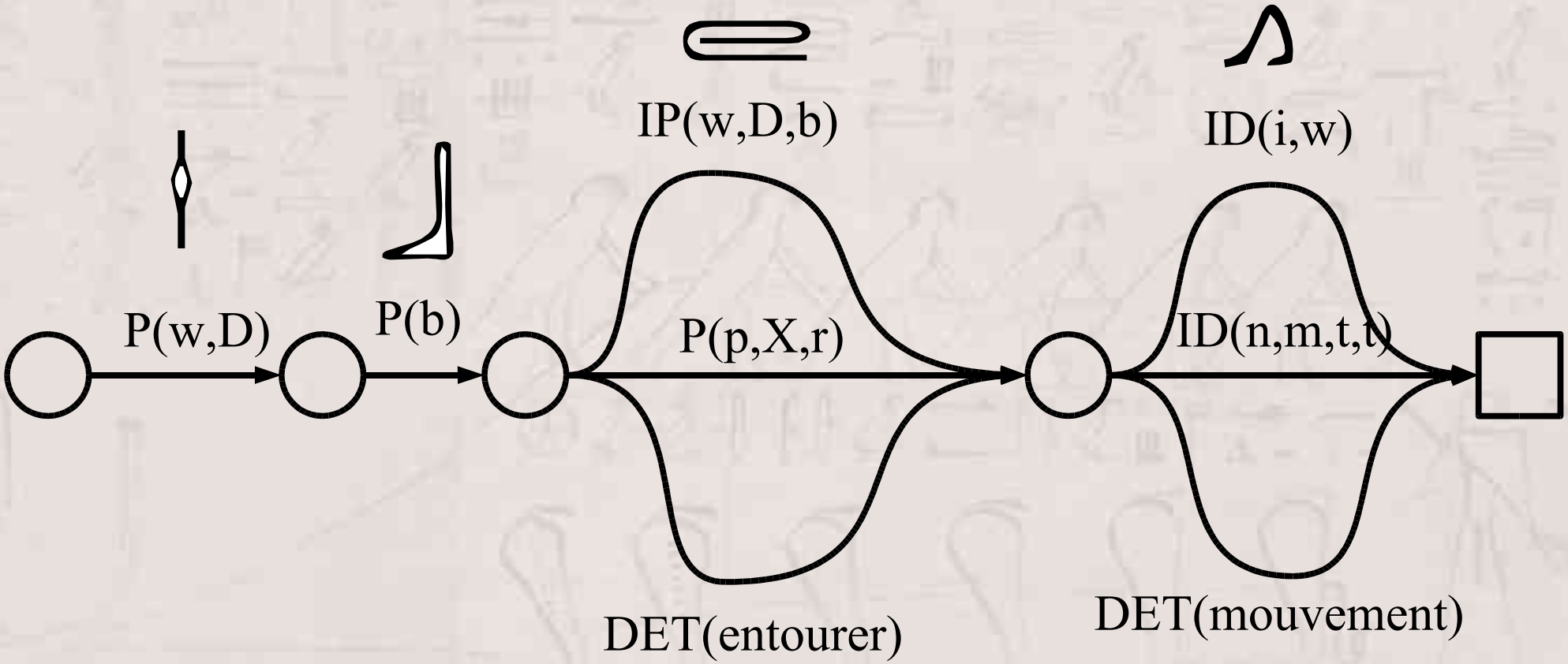
 -> I<u>D</u>(i,w)

 -> I<u>D</u>(n,m,t,t)

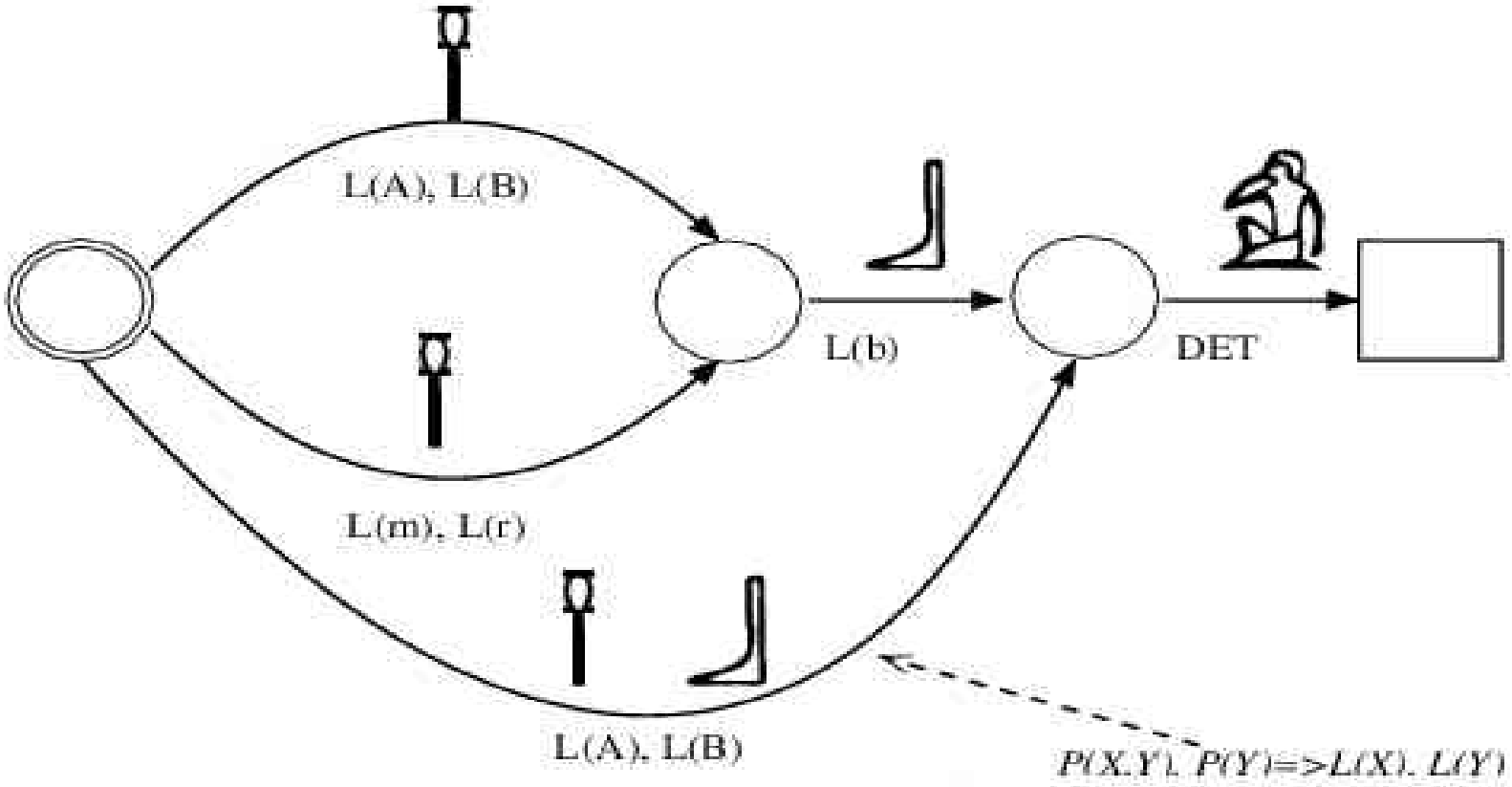
 -> DET(mouvement)

1 x 1 x 3 x 3 : 9 interprétations pour les signes

Représentation par des automates



Automates (suite)



Architecture du système

- cascade de jeux de règles
  - normalisation des signes
  - valeurs
  - groupes simples
  - formation des mots
  - déterminatifs phonétiques

Couche 1 : normalisation

- entrée : codes « manuel de codage » des signes
- sortie : codes standardisés
- les variantes de signes renvoient au signe de base

T19 => T19
qs => T19
T20 => T19
T21 => T21
wa => T21

Couche 2 : valeur des signes

- affecte à un code ses diverses valeurs possibles
- plusieurs types de valeurs :
  - P(a,n,x) : signe phonétique
  - IP(a,n,x) : idéogramme phonétique
  - ID(a,n,x) : idéogramme
  - DET(personne) : déterminatif

T14 => DET(X) / 100
T14 => P(a,m) / 100

T14 => P(n,H,s,i) / 100
T14, G41 => DET(landing) /100

Couche 3 : combinaisons de signes

- bi- et trilitères et complément phonétiques
- fins de mots

```
# qd d
P($x,$y), P($y) => G($x,$y), e2 / 10
# i+ im + m
P($x), P($x,$y), P($y) => G($x,$y), e2 / 10
```

```
# wt, yt
P(y), P(t) => b3, L(y), L(t), fin / 1000
P(w), P(t) => b3, L(w), L(t), fin / 1000
P(w), P(t), DET($x) => b3, L(w), L(t), DET($x), fin / 100
P(y), P(t), DET($x) => b3, L(y), L(t), DET($x), fin / 100
```

Couche 4 : formation de mots

- prend en compte la formation des mots
  - combinaisons des groupes précédents
  - préfixes

G($x,$y,$z) => SK(r3), L($x), L($y), L($z) / 100
G($x,$y), e2, G($z) => SK(r21), L($x), L($y), L($z) / 100
G($x,$y), e2, G($x,$y) => SK(r22), L($x), L($y), L($x), L($y) / 100

e2 => fin / 10000
e2, b3 => epsilon / 0

Quelques remarques

- En dernier ressort, les signes d'une même catégorie ont des comportements très variables : à prendre en compte pour un traitement statistique
- Extension intéressante : marquage explicite des hypothèses ;

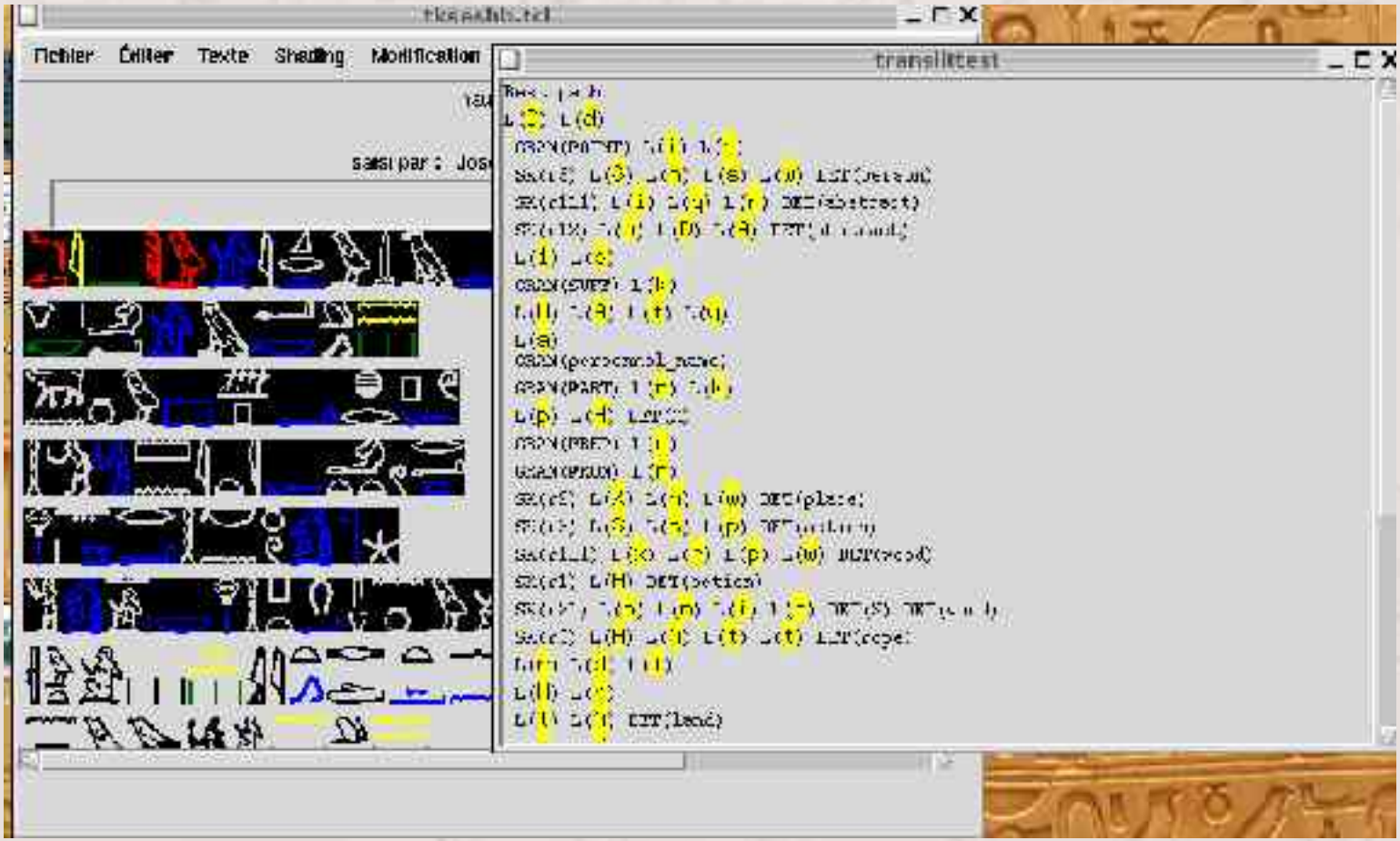
Évaluation

- Conte du naufragé :
  - 3500 signes, 1419 mots
  - 9 % d'erreurs
- Papyrus Westcar
  - 9480 signes, 4000 mots
  - 18% d'erreurs

Extensions et utilisations possibles

- Ajout d'un lexique
- normalisation de graphies
- recherche en texte intégral
- première étape vers d'autres analyses

Démo



Démonstration en ligne :

http://www.iut.univ-paris8.fr/~rosmord/Recherche/Translit/