Extraction de motifs dans des annotations de dialogues par programmation dynamique

Émilie Chanoni, Thierry Lecroq et Alexandre Pauchet
Emilie.Chanoni@univ-rouen.fr, Thierry.Lecroq@univ-rouen.fr,
Alexandre.Pauchet@insa-rouen.fr

Psychologie et Neurosciences de la Cognition et de l'Affectivité (EA 4306)
Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (EA 4108)
Université de Rouen & INSA-Rouen

Modèles Formels de l'interaction
3 – 5 juin 2009 – Lannion

Plan

1. Problématique
2. Alignement de séquences
3. Alignements de motifs 2D
4. Expérimentation
5. Conclusion et perspectives

Plan

1. Problématique
2. Alignement de séquences
3. Alignements de motifs 2D
4. Expérimentation
5. Conclusion et perspectives

Problématique

- Analyse corpus de dialogues/interactions
  - Modèle du dialogue
  - Modèle des interactions Homme-Machine
  - Modèle psychologique de l'enfant
  - Modèle psychologique de la résolution coopérative de problèmes
- Méthodologie proche
  - Recueil de corpus
  - Transcription et annotation
  - Analyse (recherche de régularités)
  - Modélisation

$\Rightarrow$ **Extraction de motifs dans des annotations par programmation dynamique**

Problématique

e25	P	don't worry	A	P	E	)	]
e26	P	*THEY ARE HIDING* themselves	A	*P*	*B*	)	]
e27	P	they are looking for	a	[	f	)	]
e28	P	who could have taken away the crown	q	[	f	)	]
e29	C	it's in ! The crown is inside the box !	a	[	f	)	]
		so *THEY* are *SUSPECTING* a lot of people	A	*P*	*Y*	C	J
e30	P	cornelius, celeste and the old lady					
e31	P	who could have stolen the crown	q	[	f	)	]
e32	C	the crown it's in !	a	[	f	)	]
e33	P	do *YOU BELIEVE*?	Q	*H*	*K*	)	]
e34	C	yes	a	[	f	)	]
e35	P	**but** B*ABAR DOESN'T KNOW* that it's in	A	*P*	*N*	**O**	**J**
e36	P	**so** *HE TELLS HIMSELF* that is a bomb, the crown	A	*P*	*N*	**C**	**J**
e37	P	or *I* don't *KNOW* what	A	*R*	*N*	)	]
e38	C	the crown	a	[	f	)	]

e8	P	where is it?	q	[	f	)	]
e10	P	*BABAR OBSERVE* cornelius give a pack to a stranger	A	*P*	*B*	)	]
e11	P	what's inside?	q	[	f	)	]
e12	P	who is this masked stranger?	q	[	f	)	]
e13	P	who has stolen the crown?	q	[	f	)	]
e14	P	*BABAR UNCOVER* the masked stranger!	A	*P*	*B*	)	]
e15	P	it's the queen celeste!	a	[	f	)	]
e16	P	*HE* is *ASKING* himself questions	A	*P*	*N*	)	]
e17	P	why the queen Celeste desguise herself?	q	[	f	)	]
e18	P	babar goes and sees the old lady to ask her about it	a	[	f	)	]
e19	C	yes	a	[	f	)	]
e20	P	*THE OLD LADY* doesn't *WANT* him to go inside!	A	*P*	*V*	)	]
e21	P	**but** just behind *HER* was *HIDDEN*	A	*P*	*B*	**O**	**J**
e22	P	a *SURPRISE* for him **in fact**	A	*P*	*S*	**O**	**J**
e23	P	and then babar goes back home	a	[	f	)	]
e24	P	and every body was here with a big gift-wrap	a	[	f	)	]

Plan

1. Problématique
2. **Alignement de séquences**
3. Alignements de motifs 2D
4. Expérimentation
5. Conclusion et perspectives

Alignement de séquences

Alignements deux à deux

- utilisés pour comparer 2 séquences x $(l = m)$ et y $(l = n)$
- comment transformer x en y ?
- largement utilisés en bioinformatique
- moyen pour visualiser les ressemblances entre 2 séquences
- basés sur des notions de distance ou de similarité
- calculés par programmation dynamique en $O(mn)$

2 types

- globaux
- locaux (algorithme de Smith et Waterman, 1981)

Alignement de séquences

Exemple

A	C	G	—	—	A
A	T	G	C	T	A

est un alignement de ACGA et ATGCTA.

Une solution peut également être donnée sous forme de script d'édition :

Opération	Séquence résultat
substitution de A par A	A
substitution de C par T	AT
substitution de G par G	ATG
insertion de C	ATGC
insertion de T	ATGCT
substitution de A par A	ATGCTA

Alignements locaux

3 opérations d'édition

- substitution d'un symbole de x à une position donnée par un symbole de y
- suppression d'un symbole de x à une position donnée
- insertion d'un symbole de y dans x à une position donnée

Scores

- $Sub(a, b)$: score de la substitution du symbole a par le symbole b
- $Del(a)$: score de la suppression du symbole a
- $Ins(a)$: score d'insertion du symbole a

Mesure de similarité

Mesure de similarité globale

$$d(x, y) = \max\{\text{score de } \gamma \mid \gamma \in \Gamma_{x,y}\}$$

où :

- $\Gamma_{x,y}$: ensemble de toutes les suites d'opérations d'édition qui transforment x en y
- le score d'un élément $\gamma \in \Gamma_{x,y}$ est la somme des scores de ses opérations d'édition élémentaires

Score d'édition

$s(x, y) =$ similarité maximale entre un segment de x et un segment de y

Programmation dynamique

$t[i,j] = s(x[0\,..\,i], y[0\,..\,j])$ pour $i = 0, \ldots, m-1$ et $j = 0, \ldots, n-1$

$s(x,y) = \max\{t[i,j]\}$

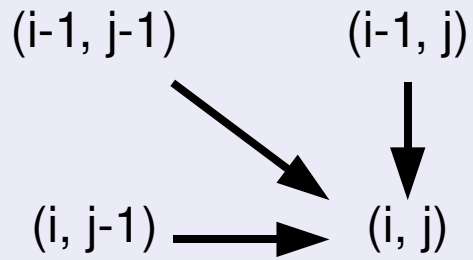
Formules de récurrence

$$
\begin{aligned}
t[-1,-1] &= 0\,, \\
t[i,-1] &= 0\,, \\
t[-1,j] &= 0\,, \\
t[i,j] &= \max \begin{cases} t[i-1,j-1] + \mathit{Sub}(x[i], y[j])\,, \\ t[i-1,j] + \mathit{Del}(x[i])\,, \\ t[i,j-1] + \mathit{Ins}(y[j])\,, \\ 0 \end{cases}
\end{aligned}
$$

pour $i = 0, 1, \ldots, m-1$ et $j = 0, 1, \ldots, n-1$

Programmation dynamique

La valeur à la position (i, j) de la table t ne dépend que des valeurs aux 3 positions voisines :



Un alignement optimal (de score maximal) peut être produit en effectuant un tracé arrière des calculs des valeurs de la table t à partir de la position maximale jusqu'à une position de valeur 0.

Alignements locaux

Exemple

$Sub(a, a) = 2$, $Sub(a, b) = -1$ et $Del(a) = Ins(a) = -2$

T	j	-1	0	1	2	3	4	5	6	7
i		$y[j]$	Z	A	T	G	C	T	A	W
-1	$x[i]$	0	0	0	0	0	0	0	0	0
0	X	0	0	0	0	0	0	0	0	0
1	A	0	0	2	0	0	0	0	2	0
2	C	0	0	0	1	0	2	2	0	1
3	G	0	0	0	0	3	1	1	0	0
4	A	0	0	2	0	1	2	0	3	1
5	Y	0	0	0	1	0	0	1	2	2

Alignements locaux

Exemple

$Sub(a, a) = 2$, $Sub(a, b) = -1$ et $Del(a) = Ins(a) = -2$

T	j	-1	0	1	2	3	4	5	6	7
i		$y[j]$	Z	A	T	G	C	T	A	W
-1	$x[i]$	0	0	0	0	0	0	0	0	0
0	X	0	0	0	0	0	0	0	0	0
1	A	0	0	2	0	0	0	0	2	0
2	C	0	0	0	1	0	2	2	0	1
3	G	0	0	0	0	3	1	1	0	0
4	A	0	0	2	0	1	2	0	3	1
5	Y	0	0	0	1	0	0	1	2	2

Alignements locaux

Exemple

$Sub(a, a) = 2$, $Sub(a, b) = -1$ et $Del(a) = Ins(a) = -2$

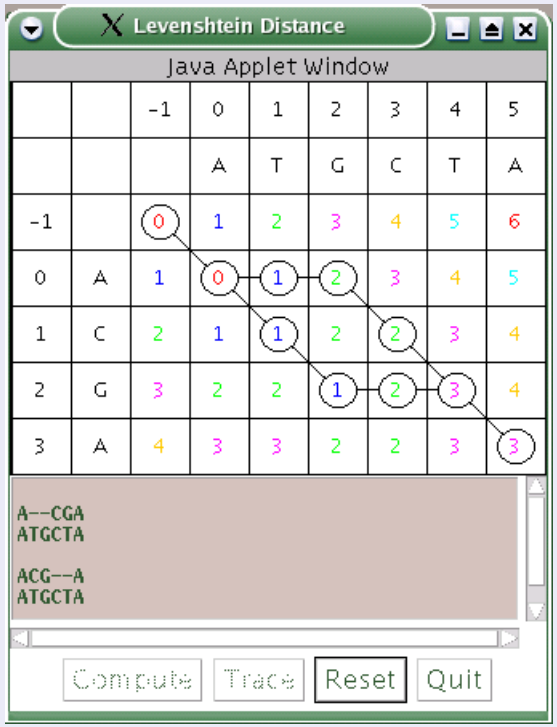
T	j	-1	0	1	2	3	4	5	6	7
i		$y[j]$	Z	A	T	G	C	T	A	W
-1	$x[i]$	0	0	0	0	0	0	0	0	0
0	X	0	0	0	0	0	0	0	0	0
1	A	0	0	2	0	**0**	0	0	2	0
2	C	0	0	0	1	0	**2**	2	0	1
3	G	0	0	0	0	3	1	**1**	0	0
4	A	0	0	2	0	1	2	0	**3**	1
5	Y	0	0	0	1	0	0	1	2	2

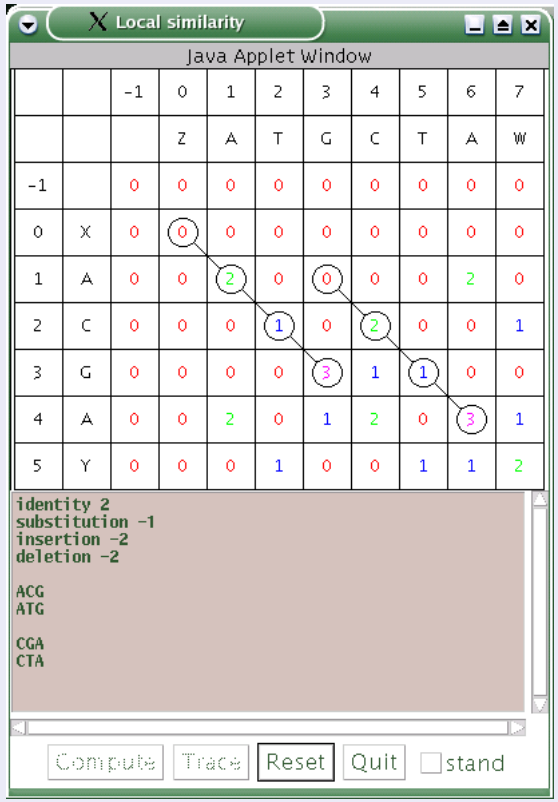
C G A
C T A

Comparaison de séquences

Sur le web

`http://monge.univ-mlv.fr/~lecroq/seqcomp`





Plan

1. Problématique
2. Alignement de séquences
3. **Alignements de motifs 2D**
4. Expérimentation
5. Conclusion et perspectives

Travaux précédents

K. Krithivasan and R. Sitalakshmi

Efficient two-dimensional pattern matching in the presence of errors

Information Sciences, 43(3), 169–184, 1987

R. Baeza-Yates

Similarity in Two-Dimensional Strings

COCOON, LNCS 1449, 319–328, 1998

A. Arslan

A Largest Common d-Dimensional Subsequence of Two d-Dimensional Strings

FCT, LNCS 4639, 40–51, 2007

2D : Alignements locaux

X	0	1	2	3	4
0	U	U	U	U	U
1	U	A	B	C	U
2	U	D	E	F	U
3	U	G	H	I	U
4	U	J	K	L	U
5	U	U	U	U	U

$|X| = m_1 \times n_1$

Y	0	1	2	3
0	V	V	V	V
1	V	E	C	V
2	V	H	I	V
3	V	K	L	V
4	V	V	V	V

$|Y| = m_2 \times n_2$

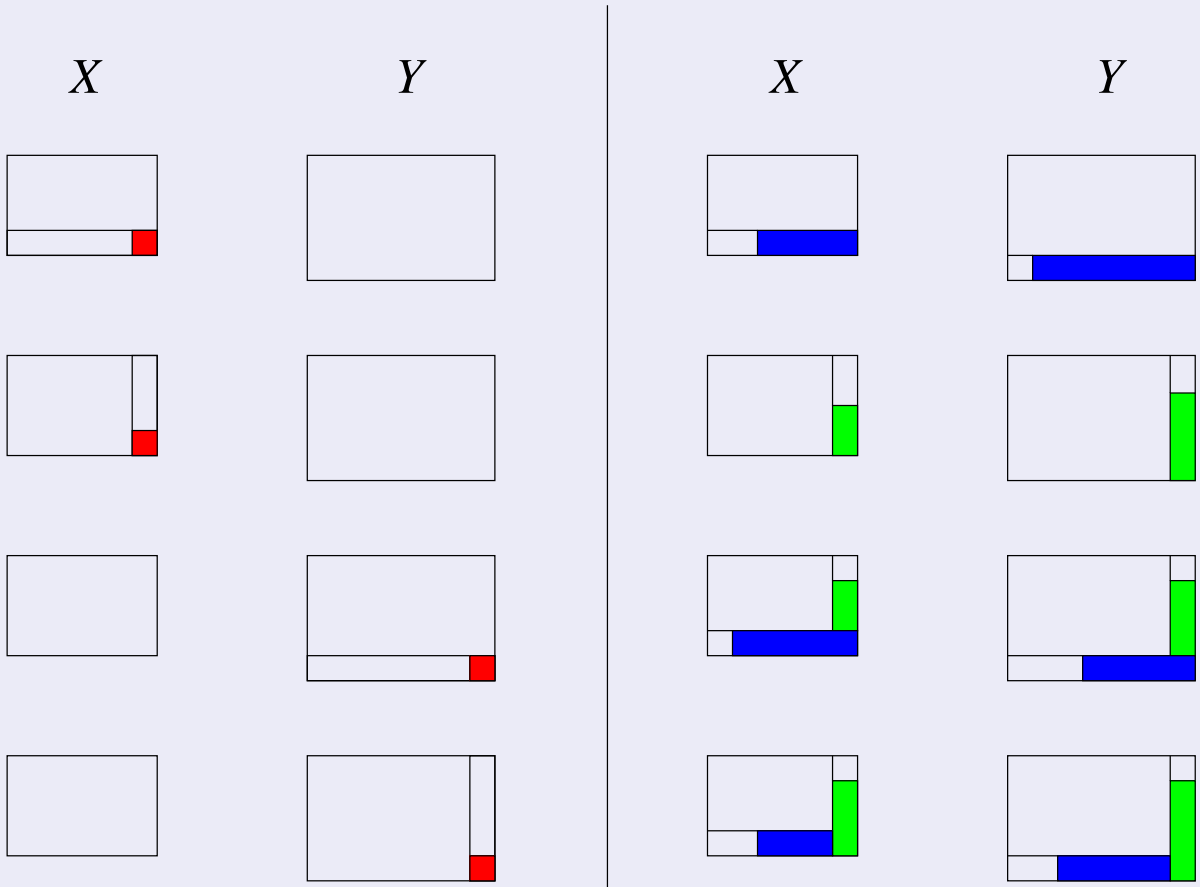
2D : Alignements locaux

2 tables de dimension 4

- $R_S[i,j,k,\ell] = s(X[i,0\,..\,j], Y[k,0\,..\,\ell])$
- $C_S[i,j,k,\ell] = s(X[0\,..\,i,j], Y[0\,..\,k,\ell])$

- $R_S[i,j,k,\ell]$: similarité maximale entre un suffixe du préfixe de longueur $j+1$ de la ligne i de X et un suffixe du préfixe de longueur $\ell+1$ de la ligne k de Y
- $C_S[i,j,k,\ell]$: similarité maximale entre un suffixe du préfixe de longueur $i+1$ de la colonne j de X et un suffixe du préfixe de longueur $k+1$ de la colonne ℓ de Y

2D : Alignements locaux



2D : Alignements locaux

$$r = R_S[i,j,k,\ell], \qquad c = C_S[i,j,k,\ell]$$

$$r' = R_S[i-1,j,k-1,\ell], \qquad c' = C_S[i,j-1,k,\ell-1]$$

Formule de récurrence

$$T[i,j,k,\ell] = \max \begin{cases} T[i-1,j,k,\ell] + \mathit{Del}[X[i,j]] \\ T[i,j-1,k,\ell] + \mathit{Del}[X[i,j]] \\ T[i,j,k-1,\ell] + \mathit{Ins}[Y[k,\ell]] \\ T[i,j,k,\ell-1] + \mathit{Ins}[Y[k,\ell]] \\ T[i-1,j,k-1,\ell] + (r \text{ si } r \neq 0 \text{ sinon } \mathit{Del}[X[i,j]] + \mathit{Ins}[Y[k,\ell]]) \\ T[i,j-1,k,\ell-1] + (c \text{ si } c \neq 0 \text{ sinon } \mathit{Del}[X[i,j]] + \mathit{Ins}[Y[k,\ell]]) \\ T[i-1,j-1,k-1,\ell-1] + (c' + r \text{ si } c', r \neq 0 \text{ sinon } \mathit{Del}[X[i,j]] + \mathit{Ins}[Y[k,\ell]]) \\ T[i-1,j-1,k-1,\ell-1] + (c + r' \text{ si } c, r' \neq 0 \text{ sinon } \mathit{Del}[X[i,j]] + \mathit{Ins}[Y[k,\ell]]) \\ 0 \end{cases}$$

2D : Alignements locaux

Exemple

X

U	U	U	U	U
U	A	B	C	U
U	D	E	F	U
U	G	H	I	U
U	J	K	L	U
U	U	U	U	U

Y

V	V	V	V
V	E	C	V
V	H	I	V
V	K	L	V
V	V	V	V

Plan

1. Problématique
2. Alignement de séquences
3. Alignements de motifs 2D
4. **Expérimentation**
5. Conclusion et perspectives

b7-BABAR

:

25 Pb7 t'inquiète pas
26 Pb7 on va la retrouver ta couronne
27 Pb7 t'inquiéte pas
28 Pb7 donc là ils se cachent
29 Pb7 ils cherchent
30 Pb7 qui pourrait avoir pris la couronne
31 b7 elle dedans, elle est dedans la couronne
32 Pb7 donc ils suspectent plein de monde, Cornélius, Céleste, la vieille dame
33 Pb7 qui a bien pu prendre la couronne ?
34 b7 la couronne elle est dedans
35 Pb7 tu crois ? !
36 b7 oui
37 Pb7 mais Babar il ne sait pas qu'elle est dedans
38 Pb7 donc il se dit que c'est une bombe, la couronne
39 Pb7 ou je ne sais quoi ?

:

b9-BABAR

:

7 Pb9 même son ami Zéphir la cherche partout avec sa loupe
8 Pb9 mais où est elle donc passée
9 Pb9 Babar remarque Cornélius donner un paquet à l'intrus
10 Pb9 Qu'est ce qui peut y avoir dedans ?
11 Pb9 mais qui est donc cet individu masqué
12 Pb9 qui a volé la couronne
13 Pb9 Babar la démasque !
14 Pb9 c'est la reine céleste !
15 Pb9 il se pose bien des questions
16 Pb9 pourquoi donc la reine Céleste s'est déguisée.
17 Pb9 Babar va donc chez la vieille dame lui demander
18 b9 oui
19 Pb9 la vieille dame ne veut pas qu'il rentre !
20 Pb9 mais derrière se cachait
21 Pb9 pour lui une surprise en réalité
22 Pb9 puis Babar rentra chez lui

.

Procédure d'analyse

- Retranscription de la totalité des dialogues
- Analyse des énoncés parentaux et enfantins
  - Analyse sémantique
  - Analyse pragmatique

Grille d'analyse

Axe sémantique

$\Rightarrow$ Référence aux états mentaux

- Emotion :"*Ils sont contents*"
- Cognition non observable : "*Ils réfléchissent*"
- Cognition observable : "*Ils espionnent*"
- Volition : "*Leo veut le râteau*"
- Epistémie : "*Il croit que c'est une bombe*"
- Hypothèse : "*C'est peut-être un bâteau*"
- Surprise : "*Babar est très étonné*"

Grille d'analyse

Axe pragmatique

- Demande d'attention (Attention générale, attention histoire)
- Type d'énoncé (Question, assertion)

- Explication/justification (Cause-conséquence, opposition, empathie)
- Contexte (contexte histoire, contexte personnel)
- Référenciation (auto-référenciation, hétéro-référenciation, référenciation au personnage)

b7-BABAR

24	a	[	{	)	]
25	A	P	E	)	]
26	A	P	B	)	]
27	a	[	{	)	]
28	q	[	{	)	]
29	a	[	{	)	]
30	A	P	Y	C	J
31	q	[	{	)	]
32	a	[	{	)	]
33	Q	H	K	)	]
34	a	[	{	)	]
35	A	P	N	O	J
36	A	P	N	C	J
37	A	R	N	)	]
38	a	[	{	)	]

b9-BABAR

8	q	[	{	)	]
9	A	P	B	)	]
10	q	[	{	)	]
11	q	[	{	)	]
12	q	[	{	)	]
13	A	P	B	)	]
14	a	[	{	)	]
15	A	P	N	)	]
16	q	[	{	)	]
17	a	[	{	)	]
18	a	[	{	)	]
19	A	P	V	)	]
20	A	P	B	O	J
21	A	P	S	O	J
22	a	[	{	)	]
23	a	[	{	)	]

Matrice de substitution

	a	q	A	Q	G	D	P	R	H	C	O	M	J	F	E	B	N	V	K	Y	S
a	10																				
q	9	10																			
A	10	9	10																		
Q	9	10	9	10																	
G	2	8	2	8	10																
D	2	8	2	8	10	10															
P	0	0	0	0	0	0	10														
R	0	0	0	0	0	0	9	10													
H	0	0	0	0	0	0	7	8	10												
C	0	0	0	0	0	0	2	2	2	10											
O	0	0	0	0	0	0	2	2	2	9	10										
M	0	0	0	0	0	0	2	8	2	8	8	10									
J	0	0	0	0	0	0	7	4	3	1	1	2	10								
F	0	0	0	0	0	0	4	7	3	1	1	7	9	10							
E	0	0	0	0	0	0	5	7	5	3	3	8	2	4	10						
B	0	0	0	0	0	0	5	5	5	3	3	3	4	4	8	10					
N	0	0	0	0	0	0	5	5	5	3	3	3	4	4	8	10	10				
V	0	0	0	0	0	0	5	5	5	3	3	3	2	4	9	7	7	10			
K	0	0	0	0	0	0	7	5	5	3	6	3	4	4	8	7	8	7	10		
Y	0	0	0	0	0	0	5	5	5	4	3	3	4	4	8	7	8	7	9	10	
S	0	0	0	0	0	0	5	5	5	3	3	6	3	4	9	7	7	7	9	8	10

Matrice de substitution

	a	q	A	Q	G	D	P	R	H	C	O	M	J	F	E	B	N	V	K	Y	S
a	10																				
q	9	10																			
A	10	9	10																		
Q	9	10	9	10																	
G	2	8	2	8	10																
D	2	8	2	8	10	10															
P	0	0	0	0	0	0	10														
R	0	0	0	0	0	0	9	10													
H	0	0	0	0	0	0	7	8	10												
C	0	0	0	0	0	0	2	2	2	10											
O	0	0	0	0	0	0	2	2	2	9	10										
M	0	0	0	0	0	0	2	8	2	8	8	10									
J	0	0	0	0	0	0	7	4	3	1	1	2	10								
F	0	0	0	0	0	0	4	7	3	1	1	7	9	10							
E	0	0	0	0	0	0	5	7	5	3	3	8	2	4	10						
B	0	0	0	0	0	0	5	5	5	3	3	3	4	4	8	10					
N	0	0	0	0	0	0	5	5	5	3	3	3	4	4	8	10	10				
V	0	0	0	0	0	0	5	5	5	3	3	3	2	4	9	7	7	10			
K	0	0	0	0	0	0	7	5	5	3	6	3	4	4	8	7	8	7	10		
Y	0	0	0	0	0	0	5	5	5	4	3	3	4	4	8	7	8	7	9	10	
S	0	0	0	0	0	0	5	5	5	3	3	6	3	4	9	7	7	7	9	8	10

b7-BABAR

24	a	[	{	)	]
25	A	P	E	)	]
26	A	P	B	)	]
27	a	[	{	)	]
28	q	[	{	)	]
29	a	[	{	)	]
30	A	P	Y	C	J
31	q	[	{	)	]
32	a	[	{	)	]
33	Q	H	K	)	]
34	a	[	{	)	]
35	A	P	N	O	J
36	A	P	N	C	J
37	A	R	N	)	]
38	a	[	{	)	]

b9-BABAR

8	q	[	{	)	]
9	A	P	B	)	]
10	q	[	{	)	]
11	q	[	{	)	]
12	q	[	{	)	]
13	A	P	B	)	]
14	a	[	{	)	]
15	A	P	N	)	]
16	q	[	{	)	]
17	a	[	{	)	]
18	a	[	{	)	]
19	A	P	V	)	]
20	A	P	B	O	J
21	A	P	S	O	J
22	a	[	{	)	]
23	a	[	{	)	]

b7-BABAR

24	a	[	{	)	]
25	A	P	E	)	]
26	A	**P**	**B**	)	]
27	a	[	{	)	]
28	q	[	{	)	]
29	a	[	{	)	]
30	A	**P**	**Y**	C	J
31	q	[	{	)	]
32	a	[	{	)	]
33	Q	**H**	**K**	)	]
34	a	[	{	)	]
35	A	**P**	**N**	**O**	**J**
36	A	**P**	**N**	**C**	**J**
37	A	**R**	**N**	)	]
38	a	[	{	)	]

b9-BABAR

8	q	[	{	)	]
9	A	**P**	**B**	)	]
10	q	[	{	)	]
11	q	[	{	)	]
12	q	[	{	)	]
13	A	**P**	**B**	)	]
14	a	[	{	)	]
15	A	**P**	**N**	)	]
16	q	[	{	)	]
17	a	[	{	)	]
18	a	[	{	)	]
19	A	**P**	**V**	)	]
20	A	**P**	**B**	**O**	**J**
21	A	**P**	**S**	**O**	**J**
22	a	[	{	)	]
23	a	[	{	)	]

b7-BABAR

24	a	[	{	)	]
25	A	P	E	)	]
26	A	P	B	)	]
27	a	[	{	)	]
28	q	[	{	)	]
29	a	[	{	)	]
30	A	P	Y	C	J
31	q	[	{	)	]
32	a	[	{	)	]
33	Q	H	K	)	]
34	a	[	{	)	]
35	A	P	N	O	J
36	A	P	N	C	J
37	A	R	N	)	]
38	a	[	{	)	]

b9-BABAR

8	q	[	{	)	]
9	A	P	B	)	]
10	q	[	{	)	]
11	q	[	{	)	]
12	q	[	{	)	]
13	A	P	B	)	]
14	a	[	{	)	]
15	A	P	N	)	]
16	q	[	{	)	]
17	a	[	{	)	]
18	a	[	{	)	]
19	A	P	V	)	]
20	A	P	B	O	J
21	A	P	S	O	J
22	a	[	{	)	]
23	a	[	{	)	]

b7-BABAR

25 Pb7 t'inquiète pas
26 Pb7 on va la retrouver ta couronne
27 Pb7 t'inquiéte pas
28 Pb7 donc là ils se cachent
29 Pb7 ils cherchent
30 Pb7 qui pourrait avoir pris la couronne
31 b7 elle dedans, elle est dedans la couronne
32 Pb7 donc ils suspectent plein de monde, Cornélius, Céleste, la vieille dame
33 Pb7 qui a bien pu prendre la couronne ?
34 b7 la couronne elle est dedans
35 Pb7 tu crois ? !
36 b7 oui
37 Pb7 mais Babar il ne sait pas qu'elle est dedans
38 Pb7 donc il se dit que c'est une bombe, la couronne
39 Pb7 ou je ne sais quoi ?

b9-BABAR

7 Pb9 même son ami Zéphir la cherche partout avec sa loupe
8 Pb9 mais où est elle donc passée
9 Pb9 Babar remarque Cornélius donner un paquet à l'intrus
10 Pb9 Qu'est ce qui peut y avoir dedans ?
11 Pb9 mais qui est donc cet individu masqué
12 Pb9 qui a volé la couronne
13 Pb9 Babar la démasque !
14 Pb9 c'est la reine céleste !
15 Pb9 il se pose bien des questions
16 Pb9 pourquoi donc la reine Céleste s'est déguisée.
17 Pb9 Babar va donc chez la vieille dame lui demander
18 b9 oui
19 Pb9 la vieille dame ne veut pas qu'il rentre !
20 Pb9 mais derrière se cachait
21 Pb9 pour lui une surprise en réalité
22 Pb9 puis Babar rentra chez lui

Plan

1. Problématique
2. Alignement de séquences
3. Alignements de motifs 2D
4. Expérimentation
5. **Conclusion et perspectives**

Conclusions

- Algorithme d'alignements de motifs 2D
  - Alignements globaux de motifs 2D
  - Alignements locaux de motifs 2D
- Application à l'analyse de dialogues

Limitations

Limitations algorithmiques

- Limitation de la taille des fichiers/dialogues
- L'algorithme implique des motifs en forme de 'L'
- Difficulté à définir un motif 'consensuel'

Limitations méthodologiques

- Visualisation des motifs 2D
- Difficulté à définir un motif 'consensuel'

Perspectives

Perspectives techniques

- Recherche des alignements dans les 4 dimensions (rotations à 90°)
- Validation statistique des motifs trouvés
- Recherche des k meilleurs alignements
- Constitution d'une banque de motifs similaires
- Visualisation des alignements en 2D

Applications futures

- Agent dialogant / narrateur
- Aide au diagnostic (situations dialogiques assymétriques)
- Formation à l'entretien
- Autres applications nécessitant la recherche de motifs en 2D