Alignement au niveau phrastique des textes parallèles français-japonais

あられ【arare】n.

1. Perle de glace. **2.** Petit biscuit de riz. **3.** **INFORM.** `AlALeR` *(système d'Alignement Autonome, Léger et Robuste)* Aligneur adapté au traitement du japonais caractérisé par l'absence d'utilisation d'analyseur morphologique et de dictionnaire.

Système AlALeR

Yayoi NAKAMURA-DELLOYE
PARIS 7 - LaTTiCe (UMR 8094)

Alignement automatique des textes parallèles

- Textes parallèles = (original + ses traductions)
- Alignement automatique
  - → Mise en correspondance des éléments des deux textes entrés
- Premières méthodes dans le cadre de travaux sur la TA
  - ✓ Méthodes basées sur la distribution lexicale (Kay & Röscheisen 93)
  - ✓ Méthodes basées sur la corrélation des longueurs (Brown et al. 91, Gale & Church 93)
  - ✓ Amélioration par l'introduction de la notion de cognat (Simard et al. 92, Langlais 97, Kraif 01)

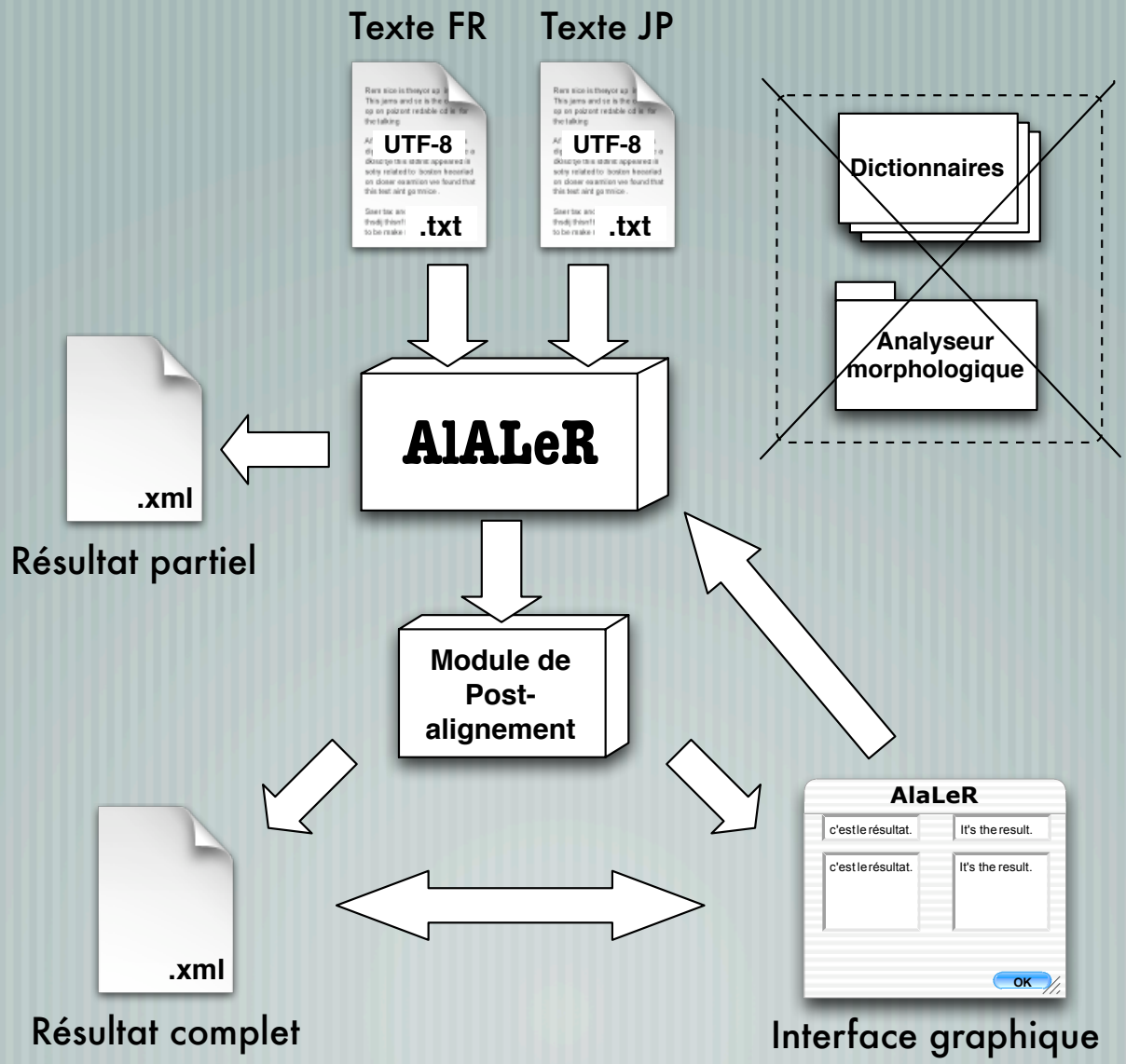
Systèmes d'alignement des textes japonais

Impossibilité d'une simple application de ces méthodes au traitement du japonais (Murao 91, Haruno & Yamazaki 96)

- ✓ Absence de séparateurs graphiques
  - ➡ Analyseur morphologique pour la segmentation
- ✓ Différences importantes sur les plans lexical et syntaxique
  - ➡ Dictionnaire bilingue

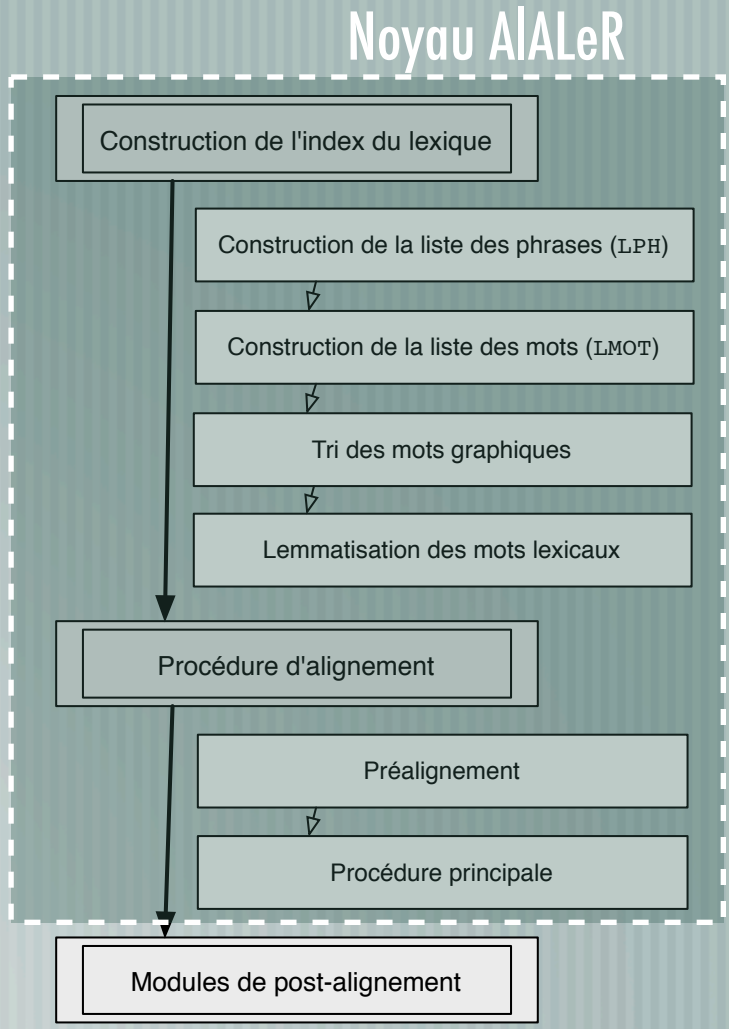
⇨ Conception d'un système autonome : AlALeR

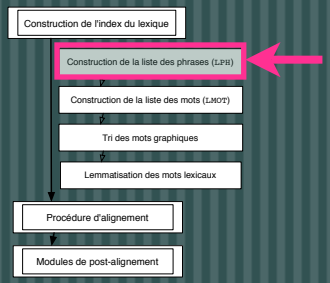
Système AlALeR



Procédure générale

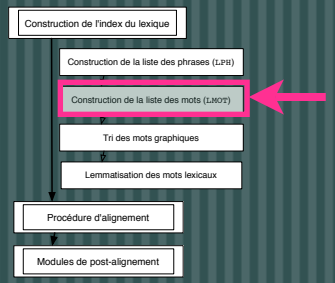
- Alignement basé sur la distribution lexicale (Kay & Röscheisen 93)
- Étape de construction de l'index du lexique
- Procédure d'alignement
- Option « Complet »
  - Module de post-alignement
  - Affichage et modification du résultat sur l'interface graphique





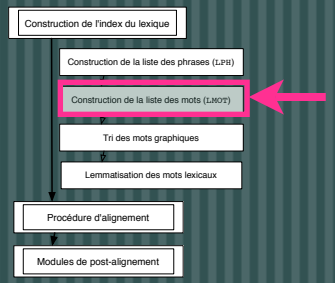
Construction de l'index du lexique (1/4)
Liste des phrases

- Repérage des points finals
- Quelques règles détaillées pour les exceptions :
  - « U.S.A. »
  - « abc@cdf.fr »



Construction de l'index du lexique (2/4)
Liste des mots

- FR : Extraction des séquences entourées de séparateurs
- JP : Analyse morphologique complète
  - ➡ Segmentation par type de caractère



Segmentation par type de caractère

Katakana

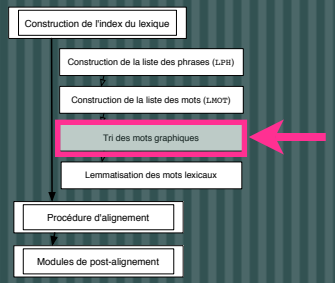
Kanji

明日 モンパルナス で 大学 の 友人 と 食事 する

Demain - Montparnasse - à - université - de - ami - avec - repas (radical) - faire (partie var.)

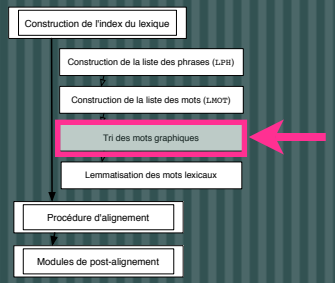
« Demain, je prendrai un repas avec des amis de l'université à Montparnasse »

Hiragana



Construction de l'index du lexique (3/4)
Tri des mots graphiques

- Liste des cognats (LCOG) → COGAL
- Liste des transfuges (LTRANS) → TRAL
- Liste des mots en katakana (LKTKN) → KTKNAL
- Liste des mots lexicaux (LEX)



Cognats, transfuges, mots en katakana

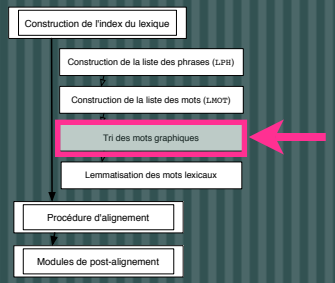
Cognats (LCOG) : mots apparentés

- generation/génération
- error/erreur

Transfuges (LTRANS) : symboles, chiffres

Mots en katakana (LKTKN) : mots emprunts

- カメラ [*ka me ra*] ← caméra
- ファイル [*fa i ru*] ← file = “fichier”



Appariement d'un mot en katakana (1/2)
Procédure générale

Table **LKTKN** (japonais)

アフリカ
コンタクト
タイプ
. . .
. . .
. . .
. . .

Retranscription

Transducteur

Listes des retranscriptions

. . .
. . .
kontakuto
kontacuto
contakuto
contacuto
. . .
. . .

Table **LMOT** (français)

. . .
. . .
considérables
consommation
contact
contre
contribuer
. . .
. . .
. . .

Tri

contact

Comparaison

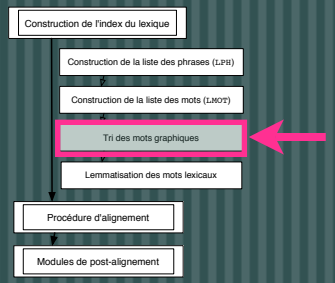
Sim(contacuto, contact) > Seuil

Listes des candidats

. . .
. . .
(contakuto, ...)
(contacuto, contact)
(contacuto, ...)
. . .
. . .
. . .

Table **KTKNAL**

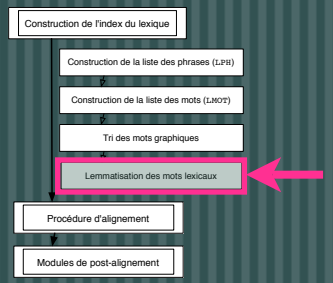
. . .
. . .
(コンタクト, contact)
. . .
. . .
. . .



Appariement d'un mot en katakana (2/2)

Transducteur et calcul de similarité

- Retranscription en une/des formes en alphabet latin par le transducteur du système
  - un AFN (89 états, 14 734 transitions)
- Détection des mots français "originaux" par calcul de similarité
  - méthode de la sous-chaîne maximale parallèle (Kraif 01)
  - adaptation aux besoins particuliers de la retranscription des katakana



Construction de l'index du lexique (4/4)
Lemmatisation des mots français

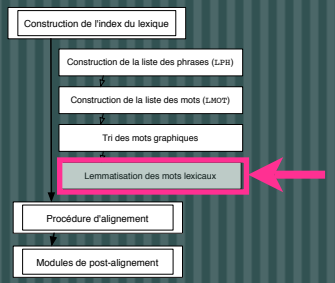
Lemmatisation sans analyseur morphologique (Kay & Röscheisen 93)

➡ Recherche des sous-chaînes communes

engagé
engager
engageons

→ "engag" + suffixe

➡ Implémentation efficace à l'aide de la structure de données **Trie** (Knuth 73)



Lemmatisation des mots japonais (1/2)

Problèmes et solution

Problèmes de la segmentation par type de caractère :

➡ Détection des frontières de mots dans une chaîne constituée d'un même type de caractère

Ex. 食糧 | 危機 | の | 原因

aliment - crise - de - origine

« l'origine de crises alimentaires »

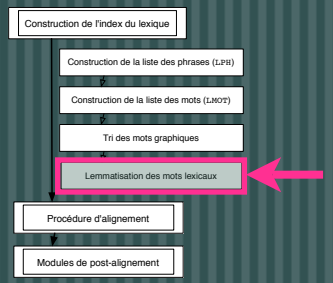
faux !

食糧危機 | の | 原因

Segmentation des mots dans une chaîne constituée d'un même type de caractère par la recherche des sous-chaînes communes

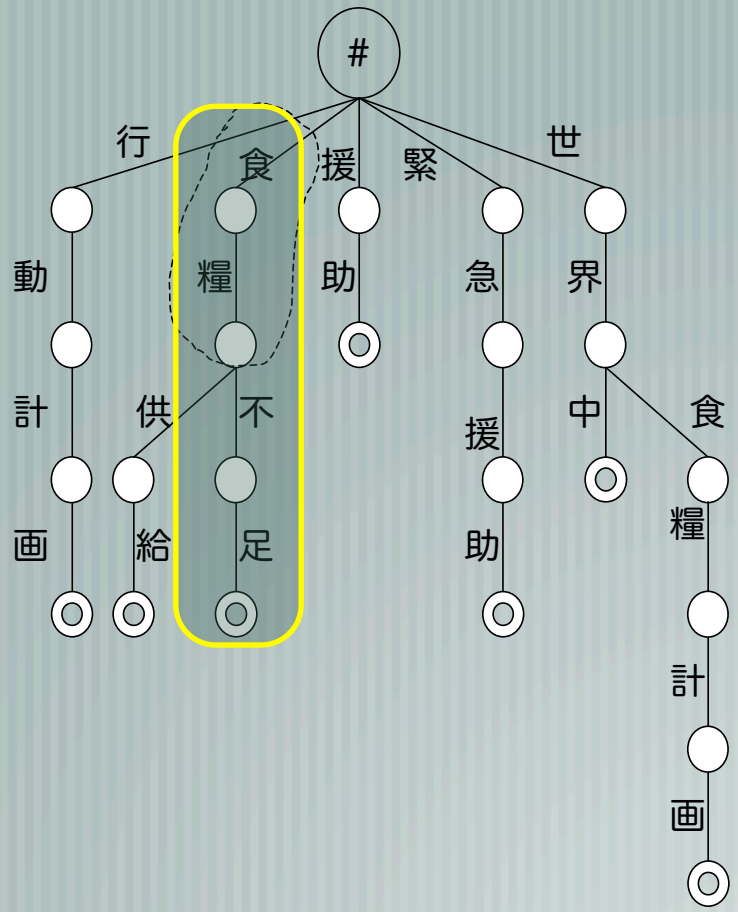
食糧供給
食糧不足
食糧危機

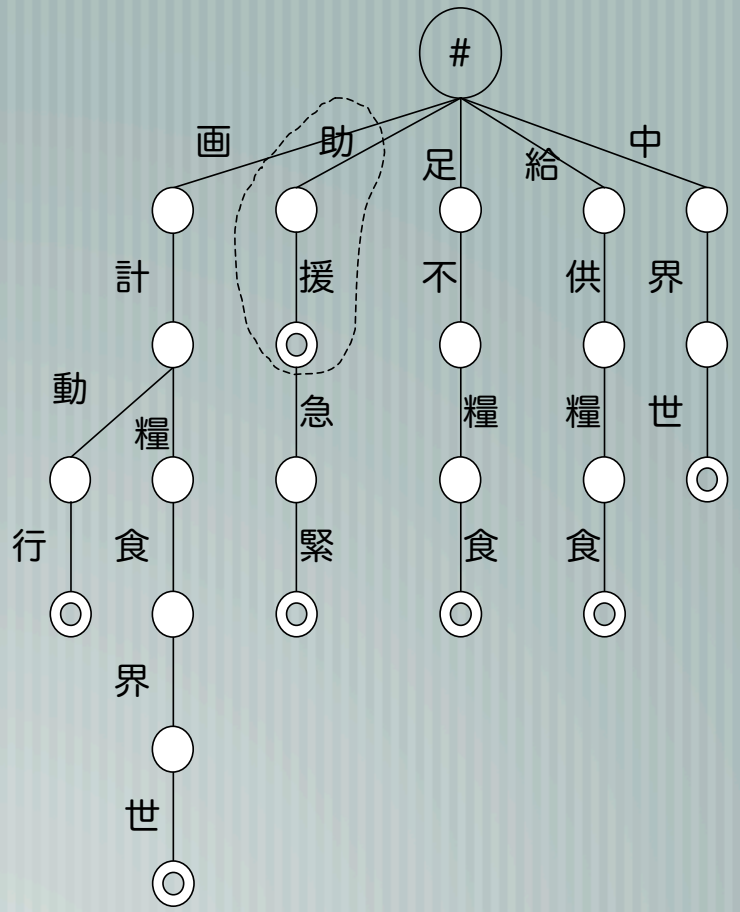
→ 食糧 ＋（供給＋不足＋危機）

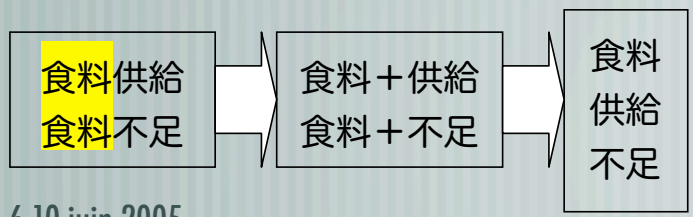


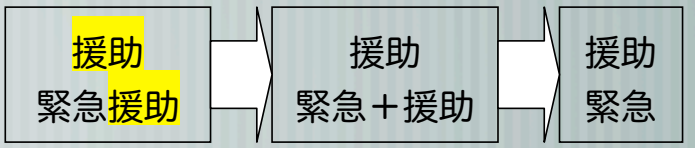
Lemmatisation des mots japonais (2/2)

Recherche des sous-chaînes communes avec Trie

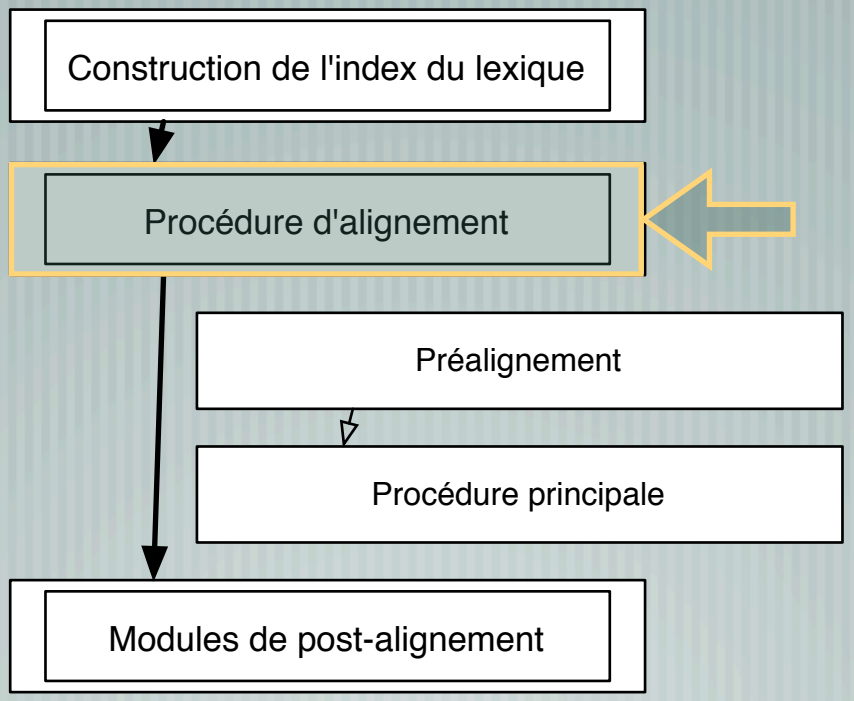


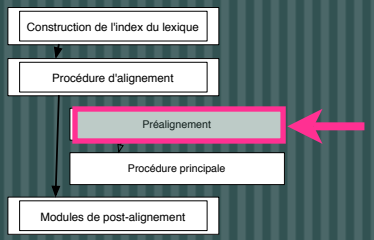






Procédure d'alignement

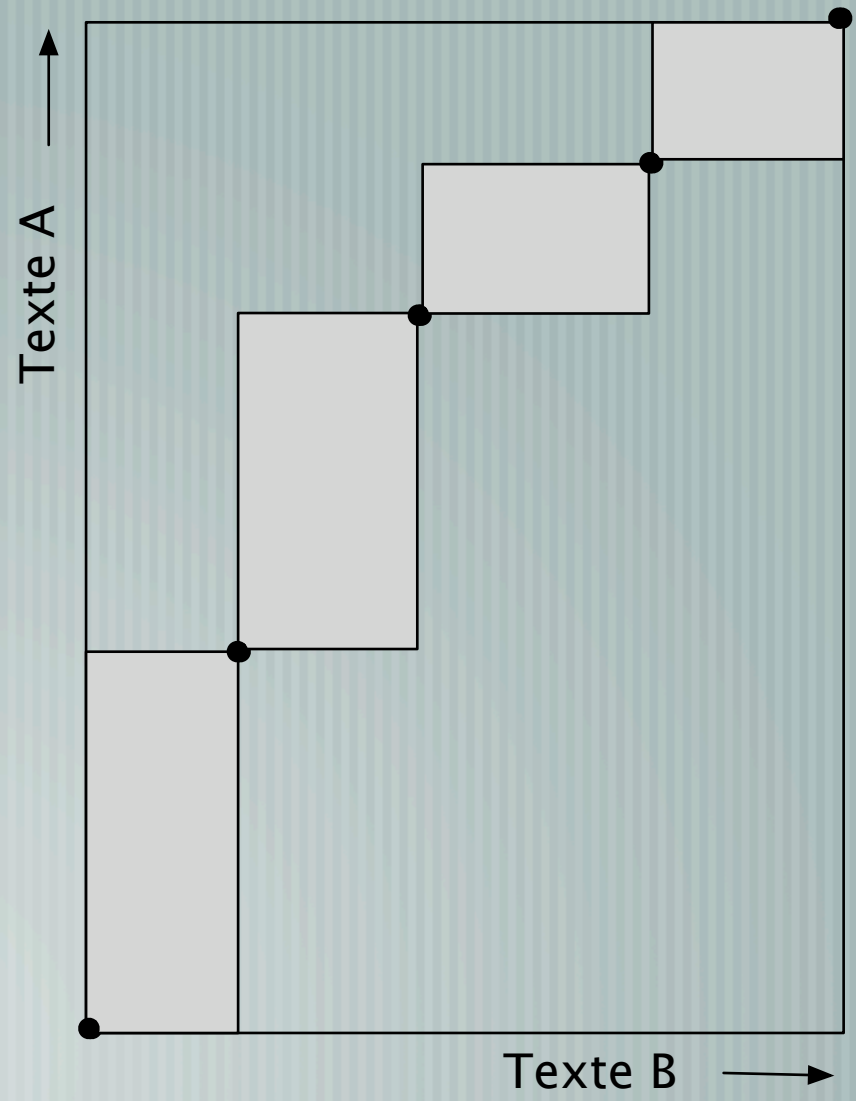


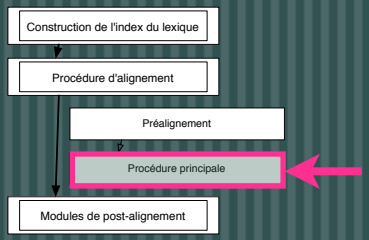


Préalignement

- ✓ Représentation par une matrice
- ✓ Préalignement = recherche des ancrages sûrs
  - ➡ réduction de la zone de recherche (Kraif 01)

Réalisation par deux parcours des tables COGAL, TRAL, KTKNAL

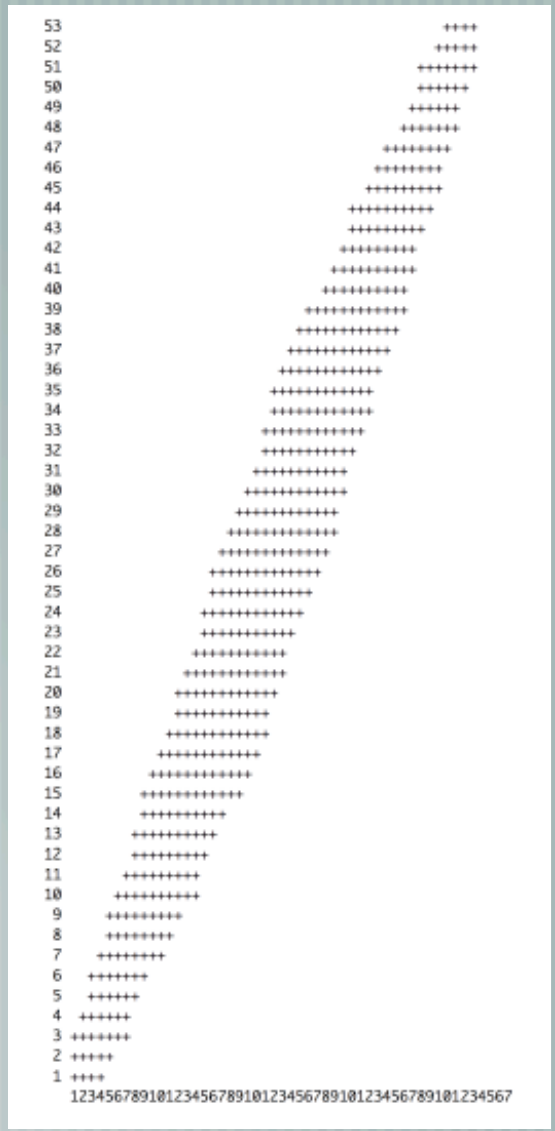


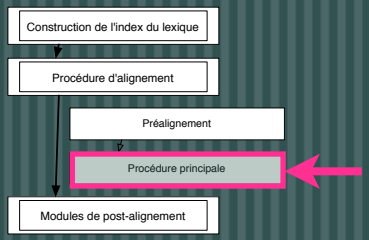


Procédure principale d'alignement

Procédure :

- ✓ Table des "Candidats paires de phrases à aligner" (CPR)
- ✓ Table des "Mots alignés" (MAL)
- ✓ Table "Résultat d'alignement" (RAL)

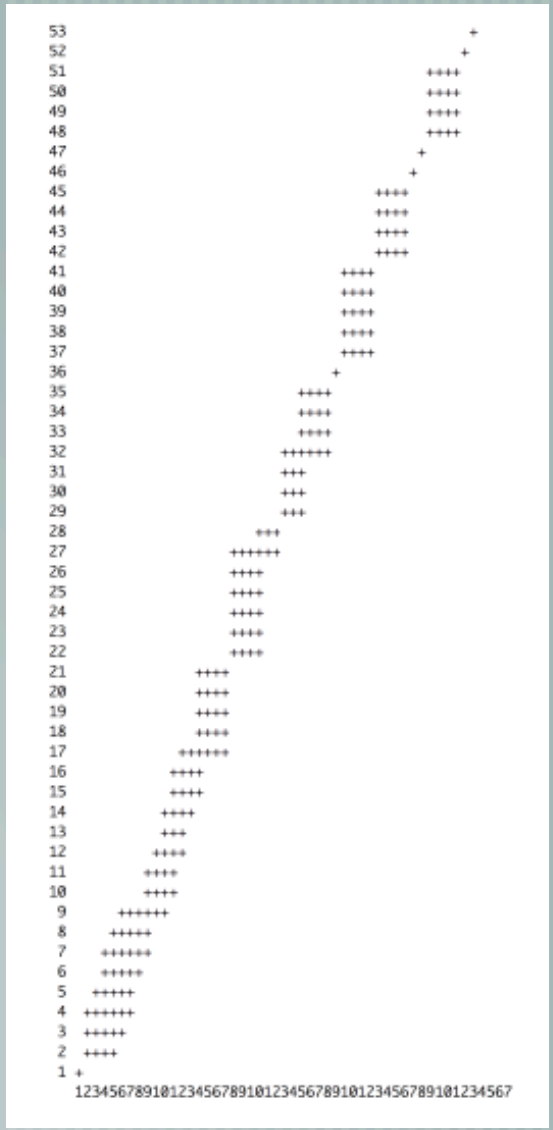


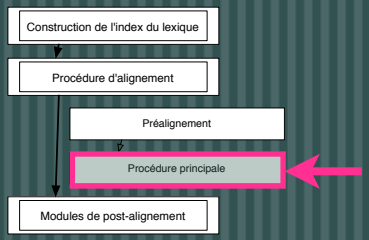


Procédure principale d'alignement

Procédure :

- ✓ Table des "Candidats paires de phrases à aligner" (CPR)
- ✓ Table des "Mots alignés" (MAL)
- ✓ Table "Résultat d'alignement" (RAL)





MAL

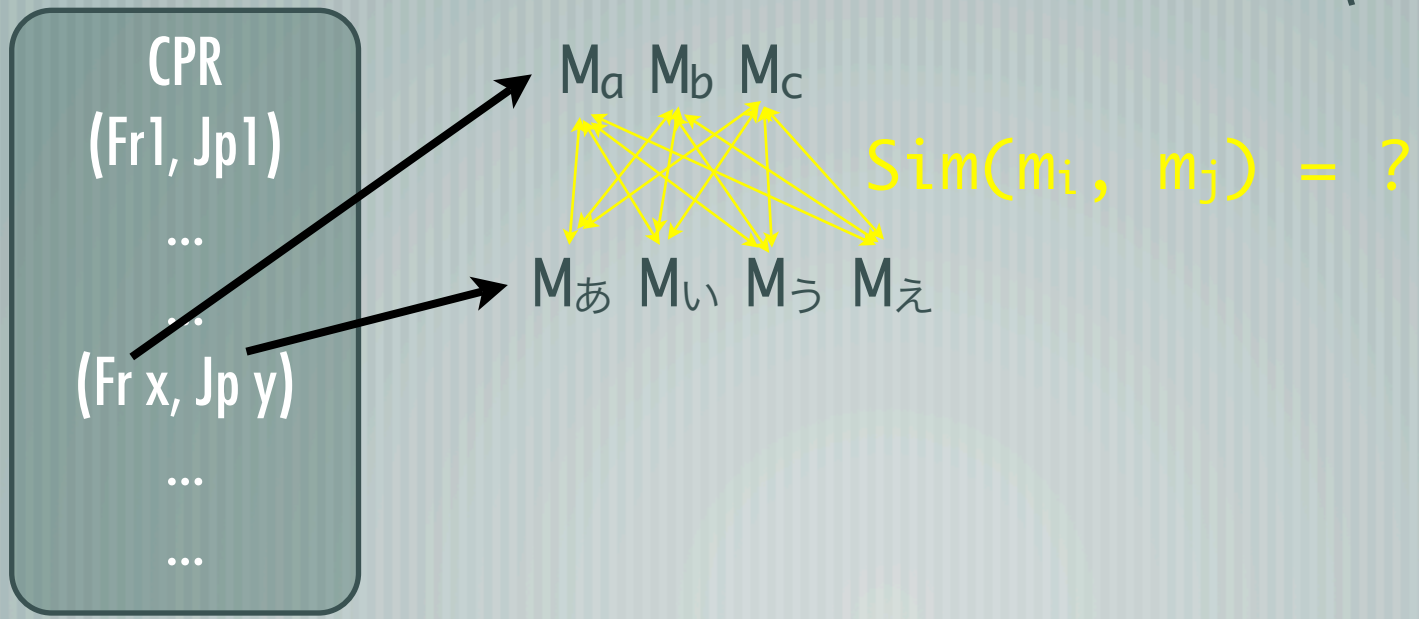
CPR → MAL → RAL

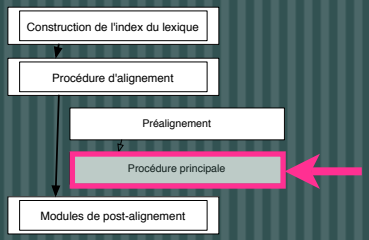
MAL :

- ✓ Comparaison de tous les mots appartenant à un même paire de phrases de la CPR
- ✓ Calcul de la similarité des mots comparés

$$sim(m_1, m_2) = p(frq(m_1, m_2)) \cdot \frac{2 \cdot frq(m_1, m_2)}{frq(m_1) + frq(m_2)} \cdot \frac{2 \cdot nb(m_1, m_2)}{nb(m_1) + nb(m_2)}$$

(Kitamura & Matsumoto 97)





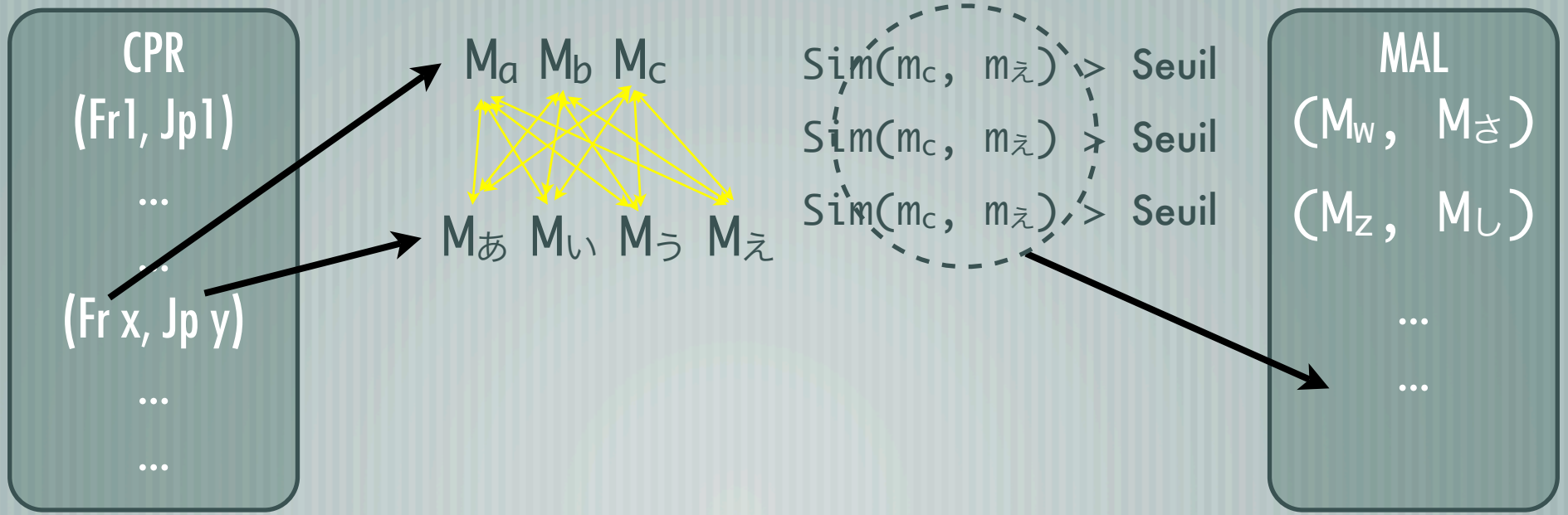
MAL

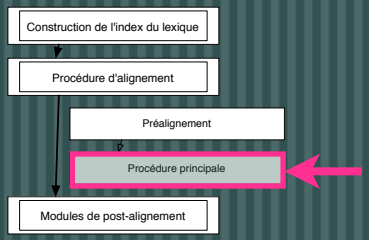
CPR → MAL → RAL

MAL :

- ✓ Comparaison de tous les mots appartenant à un même paire de phrases de la CPR
- ✓ Calcul de la similarité des mots comparés

$$sim(m_1, m_2) = p(frq(m_1, m_2)) \cdot \frac{2 \cdot frq(m_1, m_2)}{frq(m_1) + frq(m_2)}$$





MAL

CPR RAL

RAL : Paires de phrases comprenant un certain nombre de paires de mots alignés

CPR
(Fr1, Jp1)
...
...
(Fr x, Jp y)
...
...

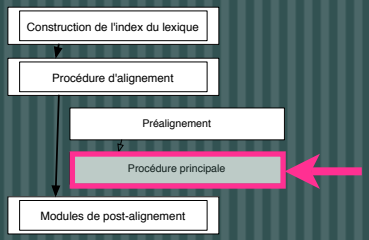
M_a M_b M_c

$M_{あ}$ $M_{い}$ $M_{う}$ $M_{え}$

MAL
$(M_a, M_{あ})(M_b, M_{い})$
$(M_c, M_{え})$ . . .

+

COGAL
TRAL
KTKNAL



MAL

CPR RAL

RAL : Paires de phrases comprenant un certain nombre de paires de mots alignés

CPR
(Fr1, Jp1)
...
...
(Fr x, Jp y)
...
...

M_a M_b M_c

$M_{あ}$ $M_{い}$ $M_{う}$ $M_{え}$

MAL
(M_a, $M_{あ}$)(M_b, $M_{い}$)
(M_c, $M_{え}$) . . .

\+

COGAL
TRAL
KTKNAL

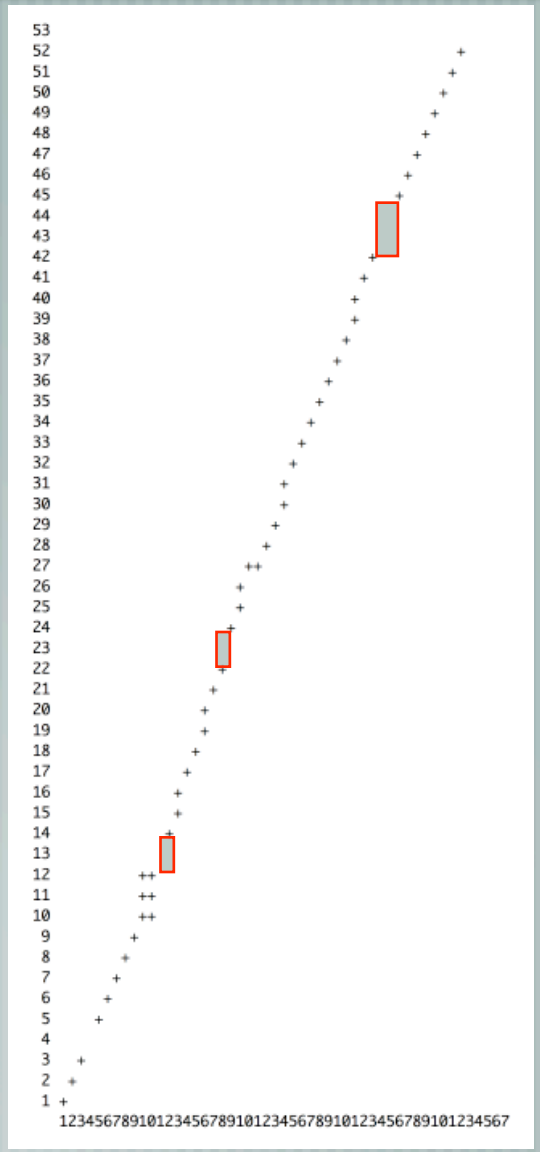
Seuil = 2

(Fr x, Jp y)

RAL
(Fr1, Jp1)
...

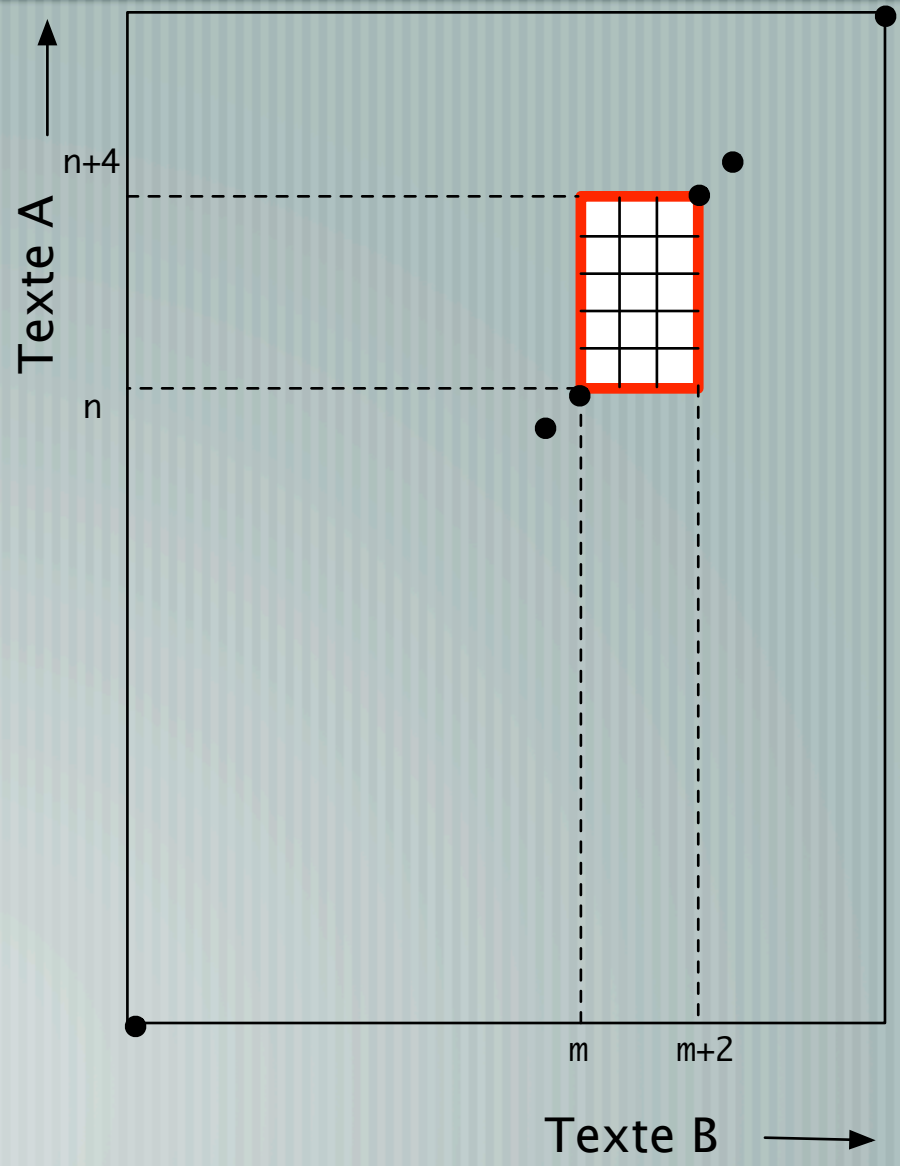
Résultat du noyau AlALeR et post-alignement

- ✓ Résultat du noyau AlALeR = alignement fiable mais partiel
- ✓ Option "Complet" : Post-alignement
  - Calcul de probabilité basé sur la longueur
  - Méthode de programmation dynamique



Résultat du noyau AlALeR et post-alignement

- ✓ Résultat du noyau AlALeR = alignement fiable mais partiel
- ✓ Option "Complet" : Post-alignement
  - Calcul de probabilité basé sur la longueur
  - Méthode de programmation dynamique

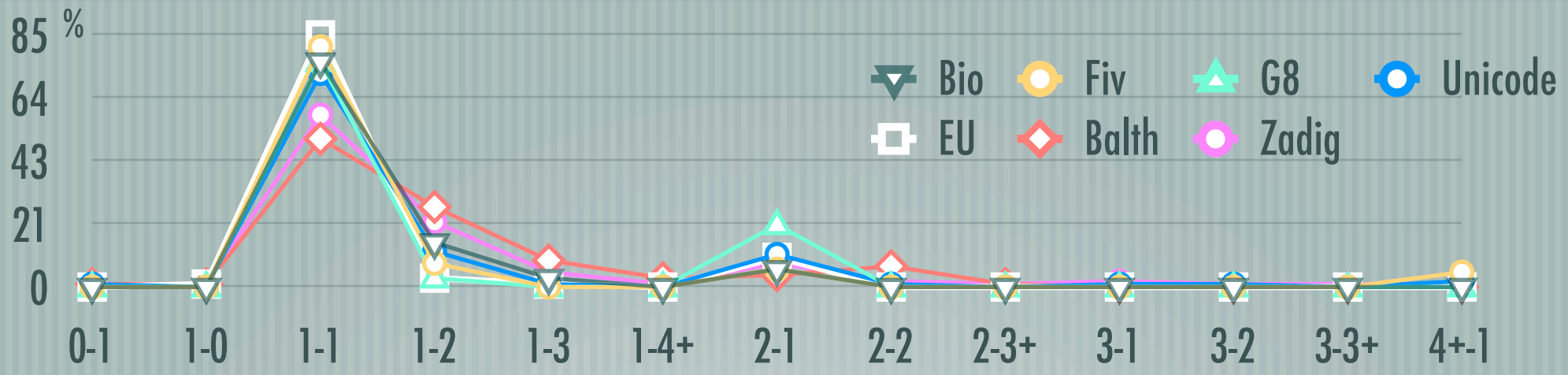


Évaluation (1/4)
Caractéristiques du corpus

	Bio		Fiv		G8		EU		Unicode		Balthasar		Zadig	
Lang	Fr	Jp	Fr	Jp	Fr	Jp	Ang	Jp	Fr	Jp	Ang	Jp	Fr	Jp
Phr	69	75	54	52	53	47	252	238	274	268	321	423	1900	2198
M/C	1 418	3 615	1 176	2 597	1 398	3 077	3 881	14 308	4 224	14 155	4 835	11 491	26 271	69 475

- **Environnement :** PowerMac G5, 512 Mo, Mac OS X 10.4, GCC 4.0
- **Textes :**
  - Bio, Fiv : deux articles de Label France sur des sujets médicaux ;
  - G8 : texte d'un sommet du G8 ;
  - EU : texte de l'Union Européenne ;
  - How to Unicode :
    - (VF) `http://www.freenix.fr/unix/linux/HOWTO/Unicode-HOWTO.html` ;
    - (VJ) `http://www.linux.or.jp/JF/JFdocs/Unicode-HOWTO.html`.
  - Balthasar, Zadig : œuvres littéraires d'Anatole France et de Voltaire.

Évaluation (2/4)
Modèles de traduction



	0 - 1	1 - 0	1 - 1	1 - 2	1 - 3	1 - 4⁺	2 - 1	2 - 2	2 - 3⁺	3 - 1	3 - 2	3 - 3⁺	4⁺ - 1
Bio	0	0	55	7	1	0	3	0	0	0	0	0	0
Fiv	0	0	43	3	0	0	2	0	0	0	0	0	1
G8	0	0	38	1	0	0	7	0	0	0	0	0	0
EU	0	4	208	5	1	0	17	0	0	0	0	0	0
Unicode	1	0	195	22	1	0	19	2	0	1	1	0	1
Balthasar	1	2	185	68	16	4	9	13	1	0	0	0	0
Zadig	7	6	1 190	300	55	9	103	20	5	18	4	1	3

Évaluation (3/4)
Résultats d'alignement

Noyau AlALeR		Bio	Fiv	G8	EU	Unicode	Balthasar	Zadig
Préalignement	Rappel	0,57	0,53	0,42	0,62	0,81	0,23	0,14
	Précision	0,98	0,93	1	0,96	0,98	0,99	0,91
Partiel	Rappel	0,81	0,66	0,95	0,87	0,90	0,49	0,66
	Précision	1	1	1	0,98	0,99	0,97	0,95
Complet	Précision	0,98	0,92	0,98	0,92	0,96	0,86	0,86

Module Post-alignement

$$\text{Rappel} = \frac{\text{Nombre de phrases alignées avec au moins une phrase de l'autre texte}}{\text{Nombre total de phrases des textes 1 et 2}}$$

$$\text{Précision} = \frac{\text{Nombre de phrases alignées correctement avec au moins une phrase de l'autre texte}}{\text{Nombre de phrases alignées avec au moins une phrase de l'autre texte}}$$

Évaluation (3/4)
Résultats d'alignement

		Bio	Fiv	G8	EU	Unicode	Balthasar	Zadig
Préalignement	Rappel	0,57	0,53	0,42	0,62	0,81	0,23	0,14
	Précision	0,98	0,93	1	0,96	0,98	0,99	0,91
Partiel	Rappel	0,81	0,66	0,95	0,87	0,90	0,49	0,66
	Précision	1	1	1	0,98	0,99	0,97	0,95
Complet	Précision	0,98	0,92	0,98	0,92	0,96	0,86	0,86

- Très bonne efficacité de l'alignement des cognats, des transfuges pour les textes informatiques
  ≠ Leur absence dans les textes littéraires (→ retours chariots, mots en katakana)
- Nombre limité de phrases alignées ← préalignement, appariement des mots
- Alignement complet robuste grâce à un alignement partiel fiable

Évaluation (4/4)
Utilisation mémoire et temps de calcul

(Zadig de 18000 mots)

Tableaux à deux dimensions

- Utilisation importante de mémoire

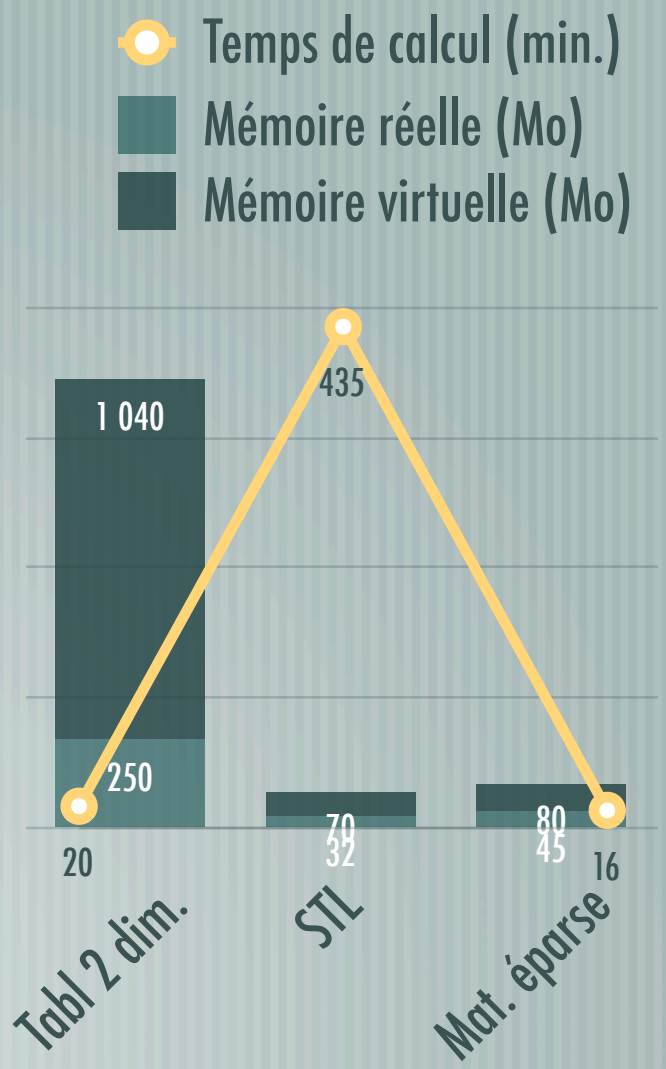
Listes de paires (STL)

`map<pair<int, int>, float>`

- Multiplication du temps de calcul

Matrices éparses

- Réduction de l'utilisation de mémoire et du temps de calcul



Conclusion

- Non utilisation d'analyseur morphologique
  - ⇒ Segmentation par type de caractère améliorée
- Non utilisation de dictionnaire
  - ⇒ Exploitation des mots emprunts
- Travaux Futurs
  - ⇒ Alignement à un niveau sous-phrastique

Merci pour votre attention