Séminaire INALCO "Corpus parallèles", 8 février 2008

Alignement automatique des propositions français-japonais

Yayoi NAKAMURA-DELLOYE

Université Paris VII, LATTICE - CNRS UMR 8094
Université Paris X, MoDyCo - CNRS UMR 7114
yayoi@free.fr

Introduction
Textes parallèles et alignement 1/2

- **Alignement** = mise en correspondance entre des unités de **textes parallèles**
- **Textes parallèles** = ensemble de textes de langues différentes, constitué d'un texte original et de ses traductions
  - Corpus compilés anglais-japonais de taille importante
  - Textes parallèles français-japonais
    - ✓ Le Monde Diplomatique
    - ✓ Label France
    - ✓ Manuels de logiciels libres
    - ✓ Textes G7, G8, etc.
    - ✓ Textes littéraires (Aozora Bunko : http://www.aozora.gr.jp/)

Introduction
Textes parallèles et alignement 2/2

- **Corpus alignés (bi- ou multitexte)**

  Différentes applications (e.g. mémoires de traduction, dictionnaires)
  - Corpus alignés anglais-japonais disponibles sur le site du NICT (National Institute of Information and Communications Technology)
  - Corpus aligné français-japonais dans le package d'évaluation ARCADE II
  - Systèmes d'alignement
    - ✓ disponibles sur Internet (ex. GIZA++)
    - ✓ Système d'alignement phrastique Fr-Jp : AlALeR (Nakamura-Delloye, 2005)
- Possibilité d'alignement à différents niveaux

Introduction
Pourquoi la proposition ? 1/2

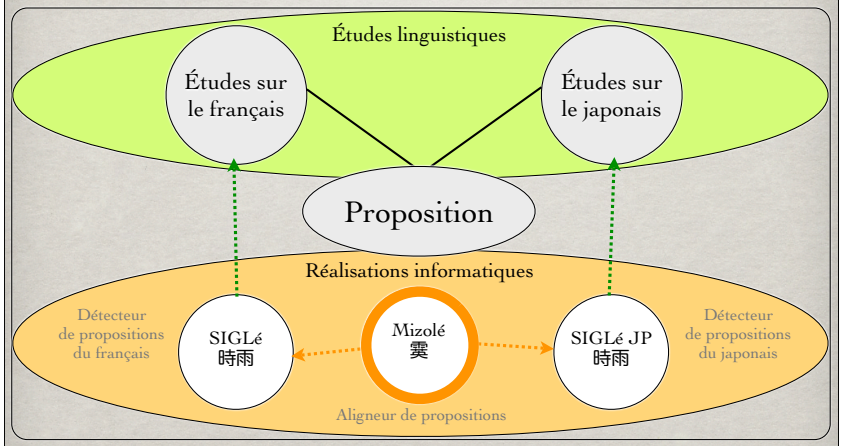
- Nécessité d'une unité de traitement plus petite que la phrase

  ➔ Proposition = bon candidat
  - la documentation technique ;
  - l'analyse discursive ;
  - ...

Introduction
Pourquoi la proposition ? 2/2

- Dans le cas de l'alignement : ex. mémoire de traduction
  - ✓ Par rapport à la **phrase** : réutilisabilité plus importante
  - ✓ Par rapport aux **unités inférieures** : portabilité plus élevée de la correspondance
    - mot français « **compte** » = 10 définitions constituées de noms japonais différents non interchangeables
    - « **tenir compte** » = « 考慮する »

Schéma des travaux

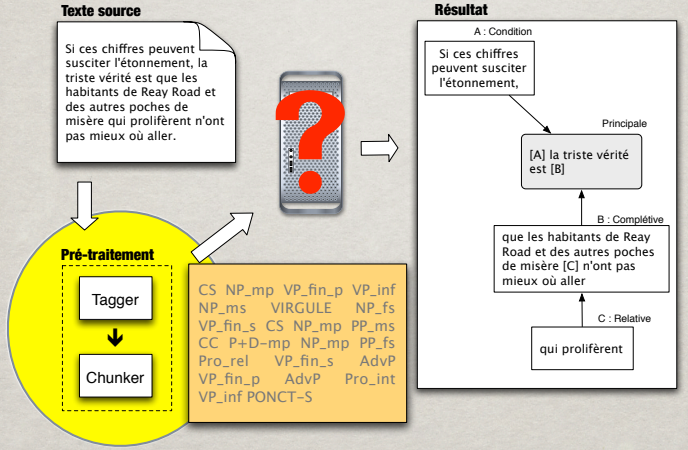


しぐれ 時雨【ʃigɯre】n.

1. Brève averse. **2.** INFORM. ***SIGLé*** *(**S**ystème d'**I**dentification de propositions avec **G**rammaire **Lé**gère)* système réalisant la détection des propositions françaises caractérisé par l'utilisation d'une grammaire hors contexte écrite dans un formalisme DCG et par une implémentation en langage PROLOG.

SIGLé FR
Système d'identification des propositions du français

Contexte
Pré-traitement



Plan

- Études linguistiques sur la proposition du français
- Réalisation
- Évaluation
- Perspectives

Études linguistiques

- **But :**
  définir une grammaire permettant de détecter des propositions à partir d'un résultat de *chunker*

➡ **Typologie des propositions, basée sur des critères formels**

➡ **Classement des subordonnées selon la position d'apparition**

Notion principale : Proposition

Proposition = sujet + prédicat

- **Types de proposition :**

1. Racine (Principale)
2. Coordonnée
   Mon père est professeur ***et ma mère travaille à la banque****.*
3. Subordonnée
   Il était déjà rentré ***quand je suis arrivé.***
4. Incidente
   Il s'en est, ***me semble-t-il****, bien sorti.*

\+ Éléments extra-prédicatifs (Charolles 1997) (Combettes 1998)
L'autre jour*, ...*
Cette affaire étant réglée*, ...*
En ce qui concerne X*, ...*

Notion principale : Proposition

Proposition = sujet + prédicat

Subordonnée

- *Il était déjà rentré* ***quand je suis arrivé.***
- *Je pense* ***qu'il viendra****.*
- *Je me demande* ***s'il est parti****.*
- *La peinture* ***qui m'a fascinée.***
- *La déception du père* ***quand il a entendu cette nouvelle.***
- *Je voterai pour* ***qui me promettra moins d'impôts****.*
- ***Où il y a de la gène****, il n'y a pas de plaisir.*
- ***Que le gouvernement propose une nouvelle loi****, l'opposition crie au scandale.*
- *Il n'a pas pu lire cette lettre* ***comme sa mère l'a deviné****.*
- *Tu peux poser ton manteau* ***où tu veux****.*
- *Je pars,* ***que cela vous plaise ou non****.*
- *Le crocodile n'eut pas le temps de se demander ce que lui voulait ce lourdaud,* ***que Gropopotin s'était déjà assis sur son dos****.*
- *La maison est restée aussi conviviale* ***qu'elle l'était avant****.*

Problèmes des typologies usuelles

- **Définitions des subordonnées souvent selon la nature du connecteur qui les introduit**

... MAIS

➡ Étiquetage automatique très difficile ;

➡ Comportements syntaxiques différents de termes appartenant à la même catégorie

➡ Qu'est-ce qu'une locution conjonctive (LC) ? Les LCs constituent-elles une liste fermée ?

Études linguistiques

Travaux de Le Goffic 1/2

(Le Goffic 1992, 1993, 2002)

- **Termes en «qu-»** = seuls connecteurs du français

  Une vieille famille indo-européenne en «*Kw-*» :
  - **pronoms** : qui, que, quoi, lequel ;
  - **adjectif** : quel ;
  - **adverbes** : où, quand, comme, comment, combien, que, dont, pourquoi.
- Subordonnées introduites par une LC = des GAdv ou GPrép comprenant une subordonnée constituée par un connecteur en «qu-»

➡ un traitement unifié et homogène des types de subordonnées

Études linguistiques

Travaux de Le Goffic 2/2

- Complétive : complétive

  *Je crois **qu'il va pleuvoir**...*
- Relative : relative avec antécédent

  *Le médecin **qui est venu** / la maison **où je suis né**...*
- Intégrative :
  - Pronominale : relative sans antécédent

    ***Qui dort** dîne / embrassez **qui vous voulez**...*
  - Adverbiale : circonstancielle en qu- ou si

    ***Quand on veut**, on peut / **si vous avez fini**, vous pouvez sortir / il est à peine sorti **qu'il a commencé à pleuvoir**...*
- Percontative : interrogative/exclamative indirecte

  *Je sais **qui a gagné** / **où il est allé** / **comment il l'a fait**...*

Études linguistiques

Autres typologies

- Typologie selon la **catégorie** du mot simple équivalent : (*Le Bon Usage*, 11ème éd.) (Biskri et Desclés 2005)
  - **Substantive :** *Je pense **qu'il viendra** / **Que tu m'aimes** me réjouit*
  - **Adjective :** *La femme **que tu vois** / la ville **où j'habite***
  - **Adverbiale :** *Il était déjà rentré **quand je suis arrivé***
- Typologie selon la **fonction** de la subordonnée dans la principale : (Chevalier *et al.* 1964) (Grevisse 1969) (Wilmet 1997)
  - **Sujet :** ***Que je sois malade** ne l'a jamais effleuré*
  - **Attribut :** *La triste vérité est **qu'il est fou***
  - **Objet :** *Marie sait **que Paul viendra***
  - **Circonstancielle :** *Il était déjà rentré **quand je suis arrivé***
  - **Complément de nom :** *la certitude **que son but était atteint***
  - **... etc.**

ÉTUDES LINGUISTIQUES

TYPOLOGIE DES SUBORDONNÉES SELON LA POSITION

- **Typologie selon la catégorie**
  - **Substantive, adverbiale, adjective**
- **Typologie selon la position**
  - **Initiale/Finale :**
    - *Il était déjà rentré* ***quand je suis arrivé***
  - **Post-verbale :**
    - *Je pense* ***qu'il*** *viendra*
  - **Autres positions SN :**
    - ***Que tu m'aimes*** *me réjouit*
  - **Post-nominale :**
    - *La femme* ***que*** *tu* ***vois*** */ la ville* ***où j'habite***
  - **Post-adj. et -adv. :**
    - *De même* ***que ...*** / bien ***que ...*** / aujourd'hui ***que ...***

➡ Description systématique de chaque type (Le Goffic, 2000)

POSITION POST-VERBALE
(SUBORDONNÉE COMPLÉMENT : subQ)

Subordonnées substantives :

- Complétives
  - *Je pense* ***qu'il viendra***.
- Intégratives pronominales (relatives sans ant.)
  - *Il a le droit d'embrasser* ***qui il veut***.
- Percontatives (ou interrogatives)
  - *Je me demande* ***s'il est parti***.
  - *Il ne m'a pas dit* ***quand il rentrerait***.
  - *Voyez* ***comme c'est facile***.

Post-verbale : Substantives

Initiale/Finale : Adverbiales

Post-nominale : Adjectives Adverbiales

Autres positions SN : Substantives

POSITION INITIALE/FINALE
(SUBORDONNÉE CIRCONSTANCIELLE : subP)

- ***Quand je suis arrivé****, il était déjà rentré.*
- ***Si tu ne manges pas****, tu ne guériras pas.*
- ***Comme elle est écrite en chinois****, il n'a pas pu lire cette lettre.*
- ***Où il y a de la gène****, il n'y a pas de plaisir.*
- ***Que le gouvernement propose une nouvelle loi****, l'opposition crie au scandale.*
- *Il n'a pas pu lire cette lettre* ***comme sa mère l'a deviné***.
- *Tu peux poser ton manteau* ***où tu veux***.
- *Je pars,* ***que cela vous plaise ou non***.
- *Le crocodile n'eut pas le temps de se demander ce que lui voulait ce lourdaud,* ***que Gropopotin s'était déjà assis sur son dos***.
- *La maison est restée aussi conviviale* ***qu'elle l'était avant***.

Post-verbale : Substantives

Initiale/Finale : Adverbiales

Post-nominale : Adjectives Adverbiales

Autres positions SN : Substantives

POSITION POST-NOMINALE
(SUBORDONNÉE DÉTERMINANTE : subR)

- Adjectives : (relatives et complétives)
  - *La peinture* ***qui m'a fascinée***
  - *La peinture* ***dans laquelle notre maison était reproduite***
  - *Ce* ***à quoi je m'attendais***
  - *L'idée* ***que tout est fini***.
- Adverbiales : (intégratives adverbiales)
  - *La déception du père* ***quand il a entendu cette nouvelle***
  - *La télévision* ***où je veux*** (publicité Orange)

Post-verbale : Substantives

Initiale/Finale : Adverbiales

Post-nominale : Adjectives Adverbiales

Autres positions SN : Substantives

Autres position SN
(subordonnée substantive : subSN)

- Position Sujet :
  - ***Qui dort*** *dîne.* (Intégrative pronominale)
  - ***Que vous avez menti*** *me déçoit.* (Complétive)
  - ***Comment il a commis ce crime*** *n'a jamais été établi.* (Percontative)
- Après une préposition :
  - *Je voterai pour* ***qui me promettra moins d'impôts.*** (Intég.pron.)
  - *Je partirai avant* ***qu'il arrive.*** (Complétive)
  - *Il faudra se poser la question de* ***pourquoi nous avons été choisis.*** (Percontative)
- Position initiale en prolepse :
  - ***Comment il a fait****, je vous le demande !* (Percontative)

Post-verbale : Substantives

Initiale/ Finale : Adverbiales

Post-nominale : Adjectives Adverbiales

Autres positions SN : Substantives

Études linguistiques
Notre typologie des connecteurs

		Int/Fin	post-V	post-N	Autres SN
Sub.	Intégrative pro.		△		△
	Percontative		✔		△
	Complétive		✔		△
Adj.				✔	
	Relative			✔	
Adv.	Intégrative adv.	✔		△	

✔ = fréquent ; △ = moins fréquent / rare

Position	post-V				post-N				Int /Fin				pos. SN			
	I	C	P	R	I	C	P	R	I	C	P	R	I	C	P	R
qui	△		✔					✔					△		△	
que		✔				✔		✔	✔					△		
dont								✔								
où			✔		?			✔	△						△	
quand			✔		△				✔						△	
comme			✔		△				✔						△	
si			✔				△		✔						△	
quoi			✔					✔							△	
lequel			✔					✔							△	
quel			✔												△	
combien			✔												△	
comment			✔												△	
pourquoi			✔												△	

I = Intégrative, C = Complétive, P = Percontative, R = Relative

Études linguistiques
- Typologie des propositions
- Classement des subordonnées

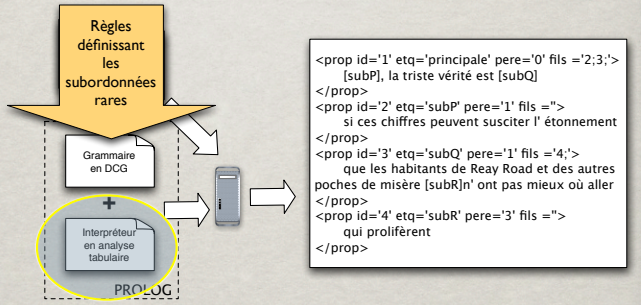
Réalisation

Évaluation

Perspectives

Réalisation

Ajout dynamique en cas d'échec de la première analyse pour chaque phrase




```
<prop id='1' etq='principale' pere='0' fils ='2;3;'>
      [subP], la triste vérité est [subQ]
</prop>
<prop id='2' etq='subP' pere='1' fils =''>
      si ces chiffres peuvent susciter l' étonnement
</prop>
<prop id='3' etq='subQ' pere='1' fils ='4;'>
      que les habitants de Reay Road et des autres
poches de misère [subR]n' ont pas mieux où aller
</prop>
<prop id='4' etq='subR' pere='3' fils =''>
      qui prolifèrent
</prop>
```

(Pereira & Shieber 1987)

Études linguistiques
- Typologie des propositions
- Classement des subordonnées

Réalisation

Évaluation

Perspectives

Évaluation 1/8

	Nb de phr	Rappel	Préc.1	Préc. 2	Préc. T
G8	53	96,2	98,0	100,0	98,0
Unicode	274	81,4	96,2	97,8	94,1
Zadig	1206	88,6	92,8	95,3	88,4
LMD*	1713	84,9	89,2	98,0	87,4

*Le Monde Diplomatique

A : Nombre total de phrases
B : Nombre de phrases dont l'analyse a abouti
C : Nombre de phrases dont les frontières de propositions sont correctement détectées
D : Nombre de phrases dont les relations des propositions sont correctement analysées

$$\text{Rappel}\ (\%) = \frac{B}{A} \times 100$$

$$\text{Précision 1}\ (\%) = \frac{C}{B} \times 100$$

$$\text{Précision 2}\ (\%) = \frac{D}{C} \times 100$$

$$\text{Précision T}\ (\%) = \frac{\text{Préc. 1} \times \text{Préc.2}}{100} = \frac{D}{B} \times 100$$

ÉVALUATION 2/8

Si ces chiffres peuvent susciter l'étonnement, la triste vérité est que les habitants de Reay Road et des autres poches de misère qui prolifèrent n'ont pas mieux où aller.

XML

```
<prop id='1' etq='racine' pere='0' fils ='2;3;'>
   [subP], la triste vérité est [subQ]
</prop>
<prop id='2' etq='subP' pere='1' fils ='">
   si ces chiffres peuvent susciter l'étonnement
</prop>
<prop id='3' etq='subQ' pere='1' fils ='4;'>
   que les habitants de Reay Road et des autres poches de misère [subR] n'ont pas mieux où aller
</prop>
<prop id='4' etq='subR' pere='3' fils ='">
   qui prolifèrent
</prop>
```


ÉVALUATION 3/8

En arrivant aux frontières qui séparent l'Arabie pétrée de la Syrie, comme il passait près d'un château assez fort, des arabes armés en sortirent.

XML

```
<prop id='1' etq='racine' pere='0' fils ='2;4;'>
   [ED] [subP], des arabes armés en sortirent
</prop>
<prop id='2' etq='ED' pere='1' fils ='3;'>
   en arrivant aux frontières [subR],
</prop>
<prop id='3' etq='subR' pere='2' fils ='">
   qui séparent l'Arabie pétrée de la Syrie
</prop>
<prop id='4' etq='subP' pere='1' fils ='">
   comme il passait près_d' un château assez fort
</prop>
```


ÉVALUATION 4/8

Tout ce qui passe sur mes terres est à moi, dit -il, aussi bien que ce que je trouve sur les terres des autres

XML

```
<prop id='1' etq='racine' pere='0' fils ='2;3;4;'>
   tout ce [subR] est à moi [Inc] aussi bien que ce [subR]
</prop>
<prop id='2' etq='subR' pere='1' fils ='">
   qui passe sur mes terres
</prop>
<prop id='3' etq='Inc' pere='1' fils ='">
   , dit –il,
</prop>
<prop id='4' etq='subR' pere='1' fils ='">
   que je trouve sur les terres des autres
</prop>
```


ÉVALUATION 5/8

- Erreurs de l'analyse linéaire :
  - étiquettes erronées attribuées par le pré-traitement ;
  - mauvaise interprétation de connecteurs :
    introducteur de proposition
    ⇔ intr. de syntagme (structure réduite sans verbe)
  - ambiguïté de la position post-prépositionnelle
    intégrative pronominale (subSN)
    ⇔ subordonnée déterminante (subR)

*Et dire [complétive 1 **qu'** au moment de son apogée, dans les années 1950, Cockerill employait encore plus de 25 000 personnes][complétive 2 **que** la ville de Seraing était toujours noire de fumée, de bruit, de monde, de travail.]*

ÉVALUATION 6/8

- Trois types d'erreurs de l'analyse structurale
  - Coordination de subordonnées sans connecteur

    De son côté, Taikong Corp. explique (1) | ***que*** *la firme n'a pas encore le droit de les vendre en France* (2) | , ***mais*** *peut les exposer* (3)
  - Ambiguïté de la virgule précédant une subordonnée : Subordonnée coordonnée ? ⇔ subR simple ?

    Personne ne m'a expliqué (1) | ***qu'****il s'agissait de la première étape de l'expansion prétendument bienveillante d'une nation nouvelle* (2) | , ***mais que*** *cette expansion signifiait en réalité l'expulsion violente des Indiens de la totalité du continent* (3) | , ***qu'****elle serait jalonnée d'atrocités indicibles* (4) | ***à l'issue desquelles*** *on parquerait les survivants dans des réserves* (5)
  - Ambiguïté des positions
    1. *C'est facile à dire* (1) | ***quand*** *on n'est pas concerné dans sa chair* (2)
    2. *Le pontife trouva dans son coeur* (1) | ***que*** *cela valait beaucoup* (2)

ÉVALUATION 7/8

(Paris avait estimé X) + (car Y) ?

Paris avait estimé (une référence aux valeurs religieuses n'était pas acceptable + car Y) ?

- Relations ambiguës

Paris avait estimé, à l'époque, (1)

‖ ***qu'****une référence aux valeurs religieuses n'était pas acceptable* (2)

‖ ***car*** *elle soulevait des problèmes politiques et constitutionnels en France.* (3)

ÉVALUATION 8/8

- Fréquence des subordonnées

occurrence(%)

		Int/Fin	post-V	post-N	Autres SN
Sub.	Intégrative pro.		△ 0	0	△ 0,4 0,8
	Percontative		✔ 2	4	△ 0 0
	Complétive		✔ 20	27	△ 0,2 0
Adj.				✔ 2	0,3
	Relative			✔ 60	57
Adv.	Intégrative adv.	✔ 15	11	△ 0	0

✔ = fréquent ; △ = moins fréquent / rare

PISTES D'AMÉLIORATION

- Amélioration du pré-traitement
  - ➡ amélioration des modules de pré-traitement
  - ➡ utilisation d'autres systèmes : système de segmentation en super-chunks (Blanc et al., 2007)
- Introduction de plus d'informations
  - ➡ risque de multiplication des calculs
- Affinement des étiquettes

  attribution d'étiquettes syntactico-sémantiques

Pistes d'amélioration

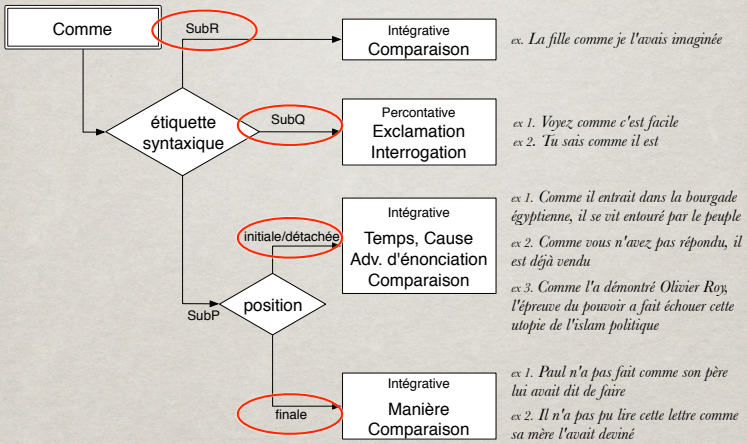
Affinement des étiquettes

Position	post-V				post-N				Int /Fin				pos. SN			
	I	C	P	R	I	C	P	R	I	C	P	R	I	C	P	R
qui	△		✔					✔					△		△	
que		✔				✔		✔	✔					△		
dont								✔								
où			✔		?			✔	△						△	
quand			✔		△				✔						△	
comme			✔		△				✔						△	
si			✔				△		✔						△	
quoi			✔					✔							△	
lequel			✔					✔							△	
quel			✔												△	
combien			✔												△	
comment			✔												△	
pourquoi			✔												△	

I = Intégrative, C = Complétive, P = Percontative, R = Relative

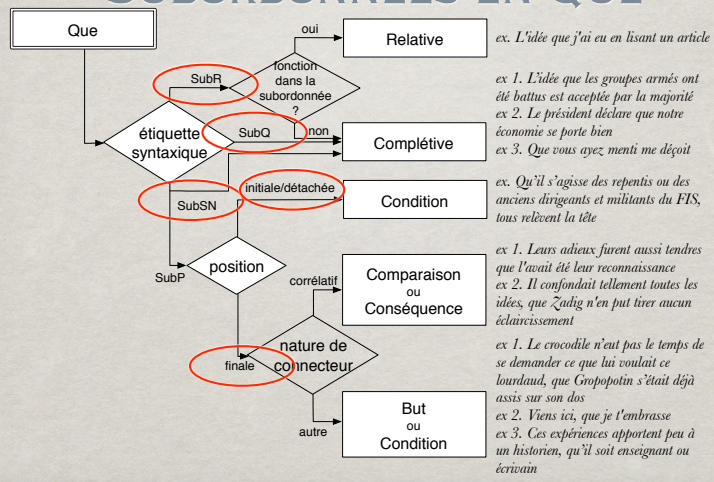
Affinement des étiquettes :

Subordonnées en Comme



Affinement des étiquettes :

Subordonnées en Que



SIGLÉ JP

Système d'identification des propositions du japonais

Plan

- Études linguistiques sur la phrase et la proposition du japonais
- Réalisation
- Évaluation

Études linguistiques

- **But :**
  définir les propositions à l'aide uniquement des critères formels pour pouvoir les identifier automatiquement

1. **Structure de la phrase japonaise, basée sur l'opposition thème-rhème**
2. **Classement des propositions japonaises**

Quelques éléments fondamentaux de la phrase japonaise

運良く	本屋 で	探していた	辞書 を	見つけた
un'yoku	*hon'ya - de*	*sagashiteita*	*jisho - wo*	*mitsuketa*
Par chance	librairie - [lieu]	que (je) cherchais	dictionnaire - [accusatif]	trouver [passé]

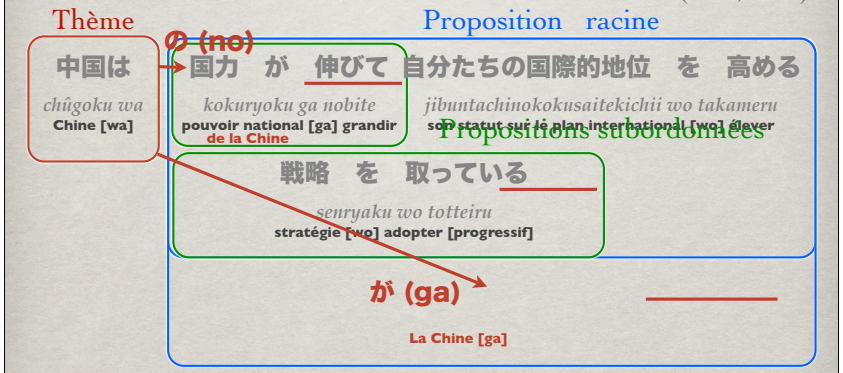
« Par chance, (j'ai) trouvé dans une librairie le dictionnaire que (je) cherchais »

- Mot prédicatif précédé par ses compléments
- Marqueurs de fonction = variations de forme, particules de cas
- Omission très fréquente d'éléments

Études linguistiques

Phrase japonaise 1/2

Mikami (1953, 1955)



« Son pouvoir national ayant grandi, la Chine adopte une stratégie permettant d'élever son statut sur le plan international »

ÉTUDES LINGUISTIQUES

PHRASE JAPONAISE 2/2

さいわいなことに 野球部員たちは、そこにあらわれたのが　たまたま　七瀬という、教務課職員とはいえ　私立手部高校随一の美人であったが為に、異変への関心をすぐ失った。　Élément externe

(*saiwaina koto ni - yakyû buin tachi wa - sokoniarawareta no ga - tamatama - nanase to iu - kyomuka shokuin towaie - shiritsu tebe kôkô zuiitsu ni bijin de atta ga tameni - ihen eno kanshin wo sugu ushinatta*)

(**par chance** - équipes du club de baseball [wa] - personne apparaissant là [ga] - par hasard - appelé Nanase - bien qu'une employée administrative - être la plus belle femme du lycée privé Tebe [cause] - perdre l'intérêt pour l'événement extraordinaire [passé])

« **Par chance**, du fait que la personne apparue était par hasard Nanase, la plus belle fille du lycée privé de Tebe, bien qu'employée administrative, les équipes du club de baseball perdirent tout de suite tout intérêt pour cet événement extraordinaire »

TYPOLOGIE DES SUBORDONNÉES

Classement selon uniquement des critères morpho-syntaxiques (Teramura 82)

- Subordonnées sans connecteur
  1. Subordonnée neutre
  2. Subordonnée de condition
  3. Subordonnée déterminante sans connecteur
- Subordonnées avec connecteur
  1. Subordonnée avec particule conjonctive
  2. Subordonnée avec connecteur agglutinant
  3. Subordonnée de citation
  4. Subordonnée déterminante avec connecteur

ÉTUDES LINGUISTIQUES

PROBLÈMES LIÉS AUX CONNECTEURS

- Subordonnée déterminante sans connecteur

日本へ　行った　友達　Substantif

nihon e - itta - tomodachi
Japon [e] - aller/partir [passé] - ami
« ami qui est parti au Japon »

- Subordonnée avec particule conjonctive

日本へ　行った　が　Particule conjonctive

nihon e - itta - ga
Japon [e] - aller/partir [passé] - [opposition]
« bien que (je sois/tu sois/il soit...) parti au Japon »

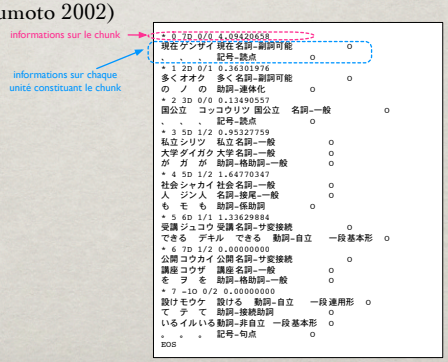
- Subordonnée avec connecteur agglutinant

日本へ　行った　時　Mot agglutinant (Sakuma 1940)

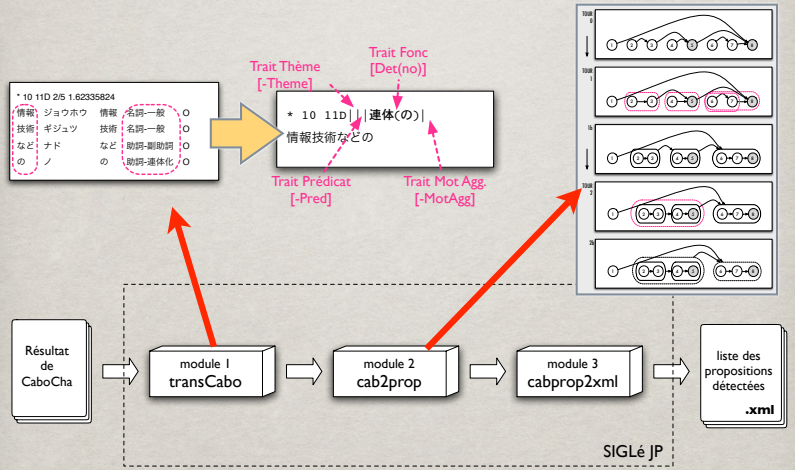
nihon e - itta - toki
Japon [e] - aller/partir [passé] - temps
« quand (je suis/tu es/il est...) parti au Japon »

RÉALISATION DU SYSTÈME

- Méthode existante : CBAP (Maruyama et al. 2004) :
  incapable de traiter les structures imbriquées
- ➡ Utilisation d'un analyseur des relations de dépendance (Kudo & Matsumoto 2002)



Réalisation du système



Études linguistiques
- Structure de la phrase
- Classement des propositions

Réalisation

Évaluation

Évaluation
Corpus

		LMD	Brevet	Asahi	Murakami
A	Nombre de phrases	124	158	112	149
B	Nombre de propositions	433	670	450	479
C	Nombre moyen de propositions dans une phrase	3,49	4,24	3,99	3,21

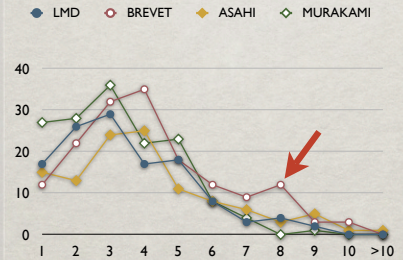


Fig. Distribution des phrases en fonction du nombre de propositions

- **LMD :** traduction d'articles d'un journal français
- **Brevet :** traduction de brevets techniques français
- **Asahi :** articles d'un journal japonais
- **FdT :** texte littéraire, extrait du roman « La fin des temps » de MURAKAMI Haruki.

Évaluation
Résultats 1/2

		LMD	Brevet	Asahi	Murakami
A	Nombre de phrases	124	158	112	149
B	Nombre de propositions	433	670	450	479
C	Nombre moyen de propositions dans une phrase	3,49	4,24	3,99	3,21
D	Nombre de propositions détectées	444	672	453	490
E	Nombre de propositions détectées **correctement**	389	589	391	426
F	Rappel (= **E/B**)	0,898	0,879	0,869	0,891
G	Précision (= **E/D**)	0,876	0,876	0,863	0,869
H	**Analyse linéaire** = nombre de phrases correctement analysées (**H/A %**)	99 (80%)	119 (75%)	85 (76%)	120 (81%)
I	**Analyse structurale** = nombre de phrases correctement analysées (**I/H %**)	94 (95%)	107 (90%)	79 (93%)	111 (93%)

Évaluation
Résultats 2/2

- Présence d'erreurs dues aux mauvaises analyses fournies par le système de pré-traitement
- Résultat des extractions expérimentales des thèmes en *wa* et des éléments externes démontrant la nécessité d'une étude linguistique plus poussée

みぞれ　霙【midzore】n.

1. grésil, neige fondue. **2.** dessert en glaçon râpé au sirop. **3.** radis blanc râpé. **4.** inform. ***MIZOLé*** système réalisant l'alignement des propositions sur la base de l'approche spectrale de l'alignement des graphes ou de la méthode inspirée de la classification ascendante hiérarchique.

MIZOLÉ

SYSTÈME D'ALIGNEMENT DES PROPOSITIONS

PLAN

- Problèmes et éléments de solution
- Trois méthodes réalisées
- Procédure d'alignement
- Évaluation
  - Constats sur les corpus utilisés
  - Résultats
  - Analyse des erreurs

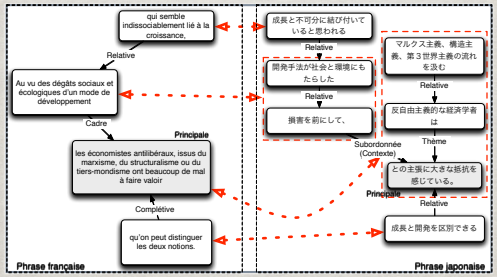
PROBLÈMES

- Peu de travaux (Boutsis & Piperidis 1998) (Wang & Ren 2005)
- Impossibilité d'une simple application des méthodes classiques d'alignement des phrases pour l'alignement fr-jp due au non-parallélisme

➡ **Alignement à l'aide des graphes**

ÉLÉMENTS DE SOLUTION

- **Alignement à l'aide des graphes**



➡ **Méthodes spectrales d'appariement des graphes**

➡ **Classification Ascendante Hiérarchique (CAH)**

Problèmes et éléments de solution

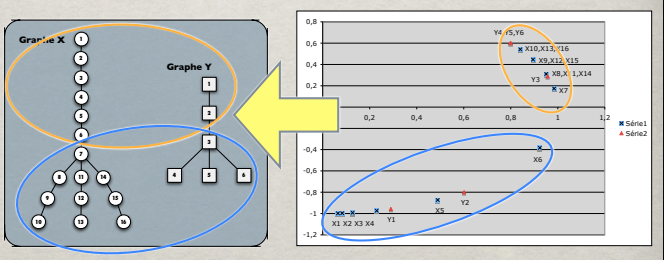
Trois méthodes

Procédure

Évaluation

Algorithme 1
Méthodes spectrales

- **Appariements des graphes inexacts (Kosinov & Caelli 2002, 2004)**
  - Combinaison des avantages des techniques de décomposition spectrale, de projection et de classification
  - Projeter les nœuds sur un sous-espace propre et regrouper les points projetés avec un algorithme de classification



Algorithme 1
Méthodes spectrales 2/2

- **Amélioration pour des graphes valués** (Lerallut 2006)

$$M^{final} = a\frac{M^{couleur}}{\max(M^{couleur})} + (1 - a)\frac{M^{topo}}{\max(M^{topo})}$$

➡ **Adaptation à l'alignement des propositions**

1. **Alignement avec la topologie (Kosinov)**
2. **Alignement des graphes valués par les types de propositions (Lerallut)**

Algorithme 2
Méthode par CAH

- **Alignement = Regroupement des points semblables**
- **Deux matrices de base**

1. Matrice de similarité :

$$M^{sim} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

$M_{11}(i,j) = synt(X_i, X_j)$
$M_{12}(i,j) = simlex(X_i, Y_j)$
$M_{21}(i,j) = simlex(X_j, Y_i)$
$M_{22}(i,j) = synt(Y_i, Y_j)$

2. Matrice d'évolution du rapport des longueurs :

$$M^{raplong}(i,j) = (rap(F(i,j), J(i,j)) - min(rap(F(i), J(i)), rap(F(j), J(j))) \cdot a$$

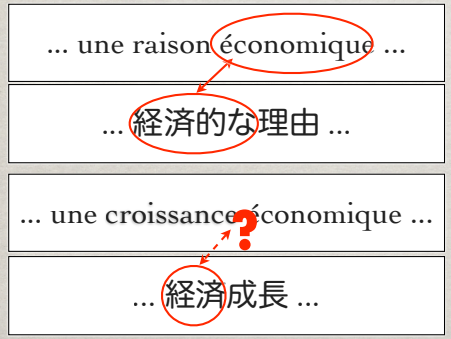
- **Matrice principale : matrice courante**

$$M^{courante}(i, j) = \frac{M^{raplong}(i, j)}{M^{sim}(i, j)}$$

Algorithme 2 : Méthode par CAH
Mise en correspondance des mots

Similarité des propositions calculée suite à la mise en correspondance des mots

➡ Inefficacité de la simple consultation du dictionnaire

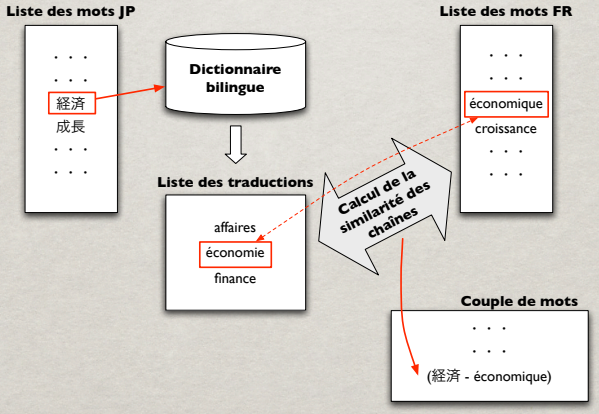


Dictionnaire

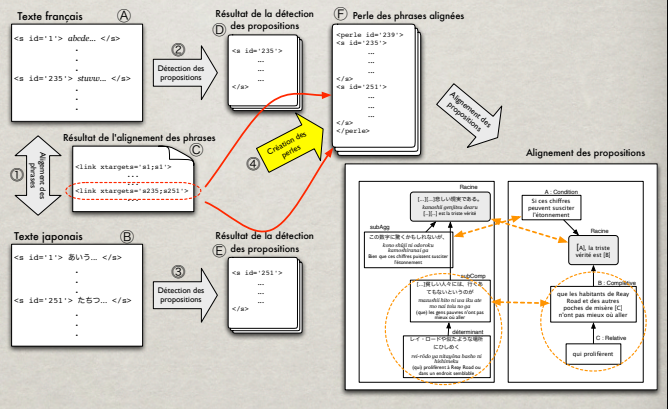
économie (n)
– 経済

économique (adj.)
– 経済的な

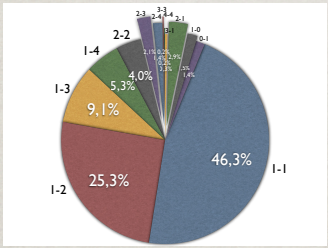
Algorithme 2 : Méthode par CAH
Mise en correspondance des mots



Procédure d'alignement



Évaluation
Corpus



	Caractéristiques			
	(A/B)	(C)	(D)	(E)
	Phr.	FR	JP	P.Prop
LMD	222/500	644	1026	583
BRVF	161/339	447	854	444
BRVJ	44/66	146	280	141
FdT	99/200	286	428	251

F → J
1. **LMD :** articles du journal LMD
2. **BRVF :** brevets techniques

J → F
3. **BRVJ :** brevets techniques
4. **FdT :** texte littéraire

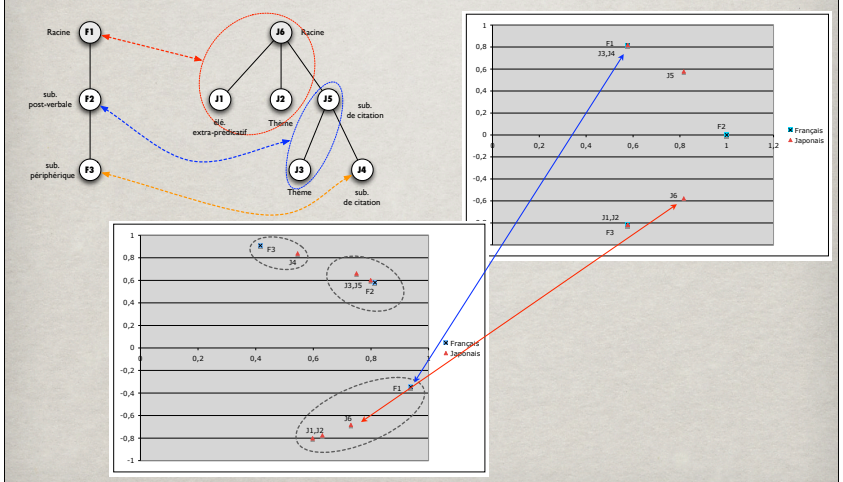
Évaluation
Résultat

Trois méthodes :
- **M1 :** Topologique
- **M2 :** Graphe valué
- **M3 :** CAH

	Exact (F)			Partiel (G)		
	M1	M2	M3	M1	M2	M3
LMD	0,127	0,200	0,591	0,643	0,784	0,951
BRVF	0,081	0,158	0,706	0,619	0,705	0,977
BRVJ	0,048	0,078	0,537	0,663	0,689	0,990
FdT	0,138	0,151	0,464	0,670	0,659	0,932

- **Résultats non satisfaisants de M1 et de M2 :** informations insuffisantes pour l'alignement

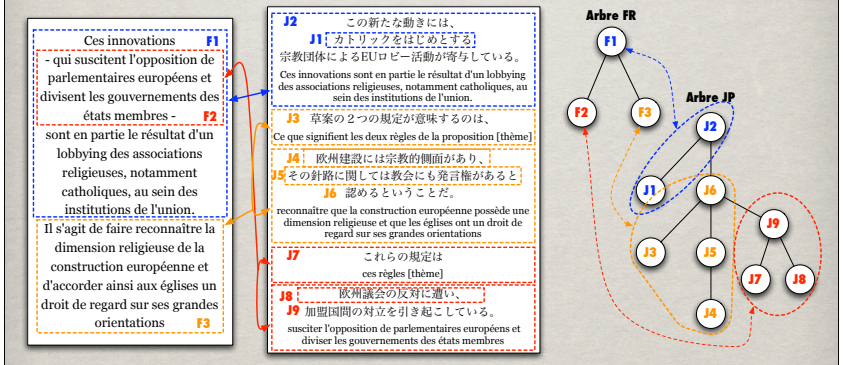
Évaluation
Problèmes des M1 et M2



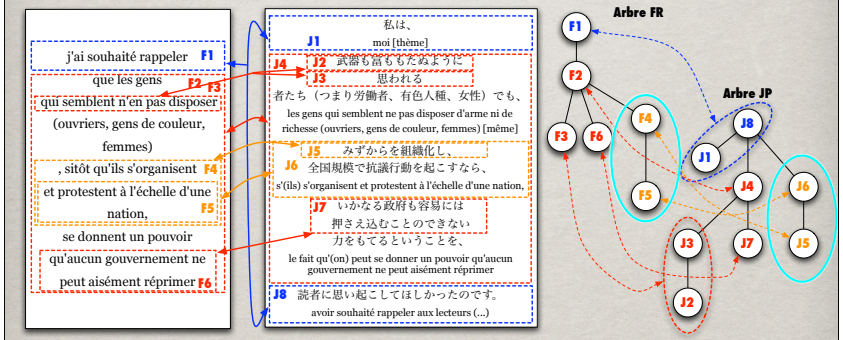
Évaluation
Méthode CAH M3

- L'introduction des informations lexicales a permis d'aligner correctement des phrases pour lesquelles la topologie et les informations sur les types des propositions ne suffisaient pas.

Évaluation
Méthode CAH M3



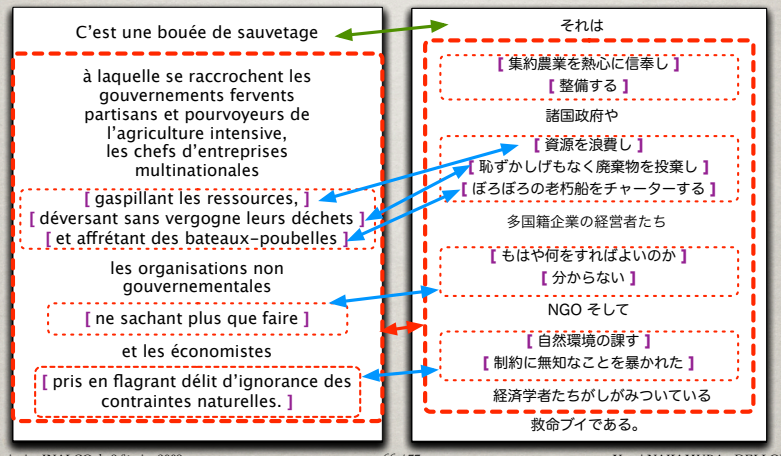
Évaluation
Méthode CAH M3



Évaluation
Méthode CAH M3

- **Méthode prometteuse mais résultat non encore tout à fait satisfaisant**
- **Mauvais résultat du calcul de similarité lexicale :**
  - ➡ Étude sur la réorganisation du dictionnaire
  - ➡ Recherche d'une meilleure méthode de mise en correspondance
- **Causes peut-être plus complexes de cette inefficacité de mise en correspondance des mots par dictionnaire**
  - ➡ À cause de l'absence d'unités dans le texte japonais ?

Évaluation
Différence du nombre des propositions



Évaluation
Cas difficiles 1/4

- Différences de diverses natures

Sur le plan lexical

Phrase française :

[F1 racine À tel point que je ne savais plus] [F2 subQ s'il progressait ou non]

Phrase japonaise :

[J1 subAgg 停まっているのか (*tomatteiru no ka*, s'il est arrêté)]
[J2 subAgg 動いているのかも (*ugoiteiru no ka mo*, s'il est en train de bouger)]
[J3 racine わからないくらいだった (*wakaranai kurai datta*, À tel point que je ne savais pas ...)]

Évaluation
Cas difficiles 2/4

Sur le plan syntaxique

Phrase française :

[F1 racine c'est notamment lors des débats sur les programmes d'aide aux pays du sud] [F2 subR que les questions de la contraception et du statut de la famille sont abordées]

Phrase japonaise :

[J1 thème 避妊と家族の地位という問題は (*hinin to kazoku no chii toiu mondai wa*, les questions de la contraception et du statut de la famille [thème])]
[J2 racine 特に開発途上国援助プログラムをめぐる議論の中で大きく取り上げられた (*tokuni kaihatsu tojô koku enjo puroguramu wo meguru giron no nakade ôkiku toriagerareta*, être abordé, notamment lors des débats sur les programmes d'aide aux pays en voie de développement)]

ÉVALUATION
CAS DIFFICILES 3/4

Sur le plan rhétorique

Phrase française :

[F1 racine Je n'arrivais pas à croire] [F2 subQ que c'était moi] [F3 subR qui avais émis ce bruit]

Phrase japonaise :

[J1 racine 私には (*watashi ni wa*, à moi)[J2 subCit それが (*sore ga*, ceci [ga]) [J3 subR 自分の体から発せられた (*jibun no karada kara hasserareta*, émis de mon corps)] 音だとは (*oto da to wa*, être un son/bruit [citation+wa])]どうしても思えなかった (*dôshitemo omoenakatta*, je n'arrivais pas à croire...)]

Je n'arrivais pas à croire que c'était un bruit émis de mon corps.

ÉVALUATION
CAS DIFFICILES 4/4

Sur le plan rhétorique

Phrase française :

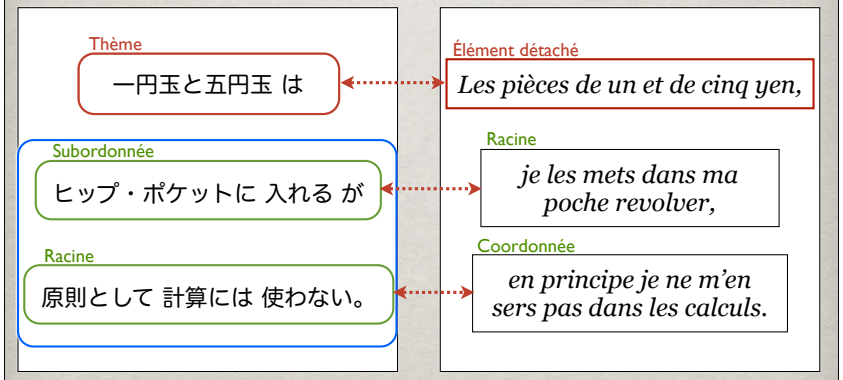
[F1 ED En y réfléchissant,] [F2 ED les trucages,] [F3 racine je n'étais pas près de les découvrir :] [F4 ED déjà,] [F5 propcrd je ne savais pas] [F6 subQ si l'ascenseur marchait ou non]

Phrase japonaise :

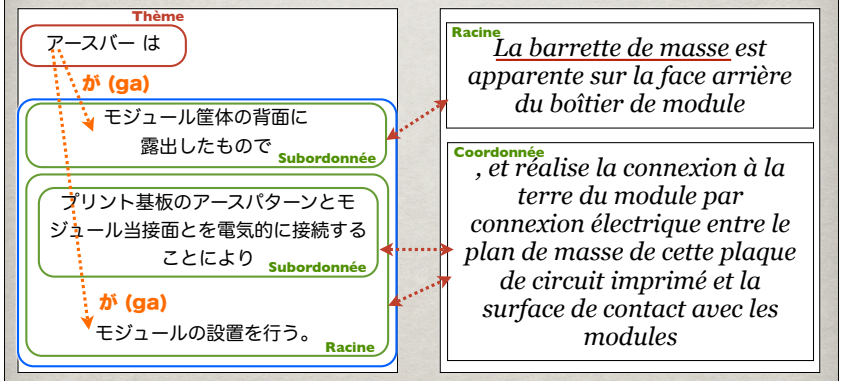
[J1 subCond 考えてみれば (*kangaete mireba*, si (je) réfléchis)] [J2 subCit たねどころか私には (*tane dokoro ka watashi ni wa*, sans aller jusqu'aux trucages, à moi) [J3 subAgg エレベーターが動いているのか (*erebêtâ ga ugoiteiru no ka*, si l'ascenseur est en train de bouger)] [J4 subAgg 停まっているのかさえ (*tomatteiru no ka sae*, s'il est arrêté)] わからないのだ (*wakaranai no da*, je ne sais pas...)]

Si je réfléchis bien, je ne sais, sans aller jusqu'aux trucages, même pas si l'ascenseur est en train de bouger ou s'il est arrêté

ÉVALUATION
ALIGNEMENT DU SYNTAGME THÉMATISÉ



ÉVALUATION
ALIGNEMENT DU SYNTAGME THÉMATISÉ



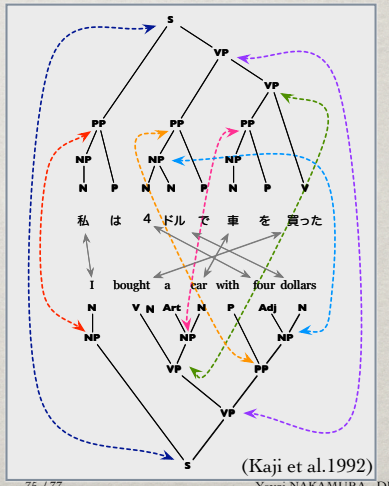
Conclusion

Conclusion & Perspectives 1/2

- Indépendance de chaque système : différentes pistes d'amélioration pour chacun
- Amélioration du concept même du système d'alignement des propositions
- Ré-examen des unités à aligner (alignement hiérarchique)

Conclusion & Perspectives
Alignement Hiérarchique

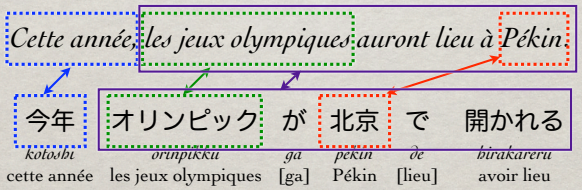
- Mise en correspondance d'unités de tout niveau sans identification préalable d'un type d'unité déterminé



(Kaji et al.1992)

Conclusion & Perspectives
Alignement Hiérarchique

- Mise en correspondance de structures
  - ➡ Liste des patrons parallèles



Conclusion & Perspectives 2/2

- Fort investissement dans les études linguistiques, mais un grand nombre de questions en suspens dans les travaux linguistiques du japonais

➡ Problèmes liés au traitement des syntagmes thématisés

Problèmes liés au syntagme thématisé

一円玉と五円玉 は
les pièces de un et de cinq yen - [wa]

を (wo)

ヒップ・ポケットに 入れる が
dans la poche revolver - mettre - [opposit.]

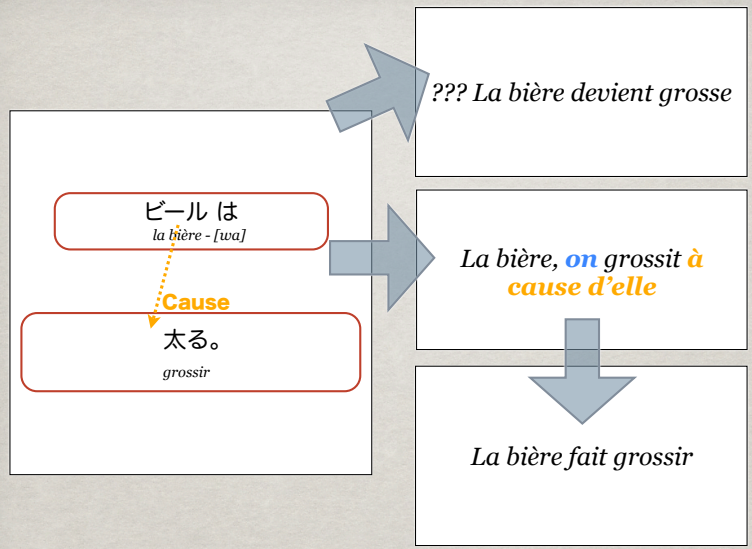
を (wo)

原則として 計算には 使わない。
en principe - dans les calculs - utiliser [nég.]

Les pièces de un et de cinq yen,

je les mets dans ma poche revolver,

en principe je ne m'en sers pas dans les calculs.



Merci