Détection des propositions syntaxiques du français en vue de l'alignement des propositions de textes parallèles français-japonais

Yayoi Nakamura-Delloye

Université Denis Diderot - Paris 7, Lattice
yayoi@free.fr

Résumé

Nous présentons dans cet article SIGLé (Système d'Identification de propositions avec Grammaire Légère), un système réalisant la détection des propositions françaises. Ce système détecte les propositions – à partir de phrases en entrée ségmentées et étiquetées en *chunk* par un analyseur extérieur –, analyse leurs relations et leur attribue une étiquette indiquant leur nature syntaxique. Il est caractérisé d'une part par sa grammaire de type CFG proposant un ensemble d'étiquettes adaptées à notre analyse pour les mots dits en « qu- », et d'autre part par l'utilisation du formalisme DCG et du langage PROLOG.

Mots-clés : analyse syntaxique partielle, proposition syntaxique, subordination, Prolog, CFG, DCG.

Abstract

In this paper we present SIGLé (Clauses Identification System with Light Grammar), which is a system recognizing French clauses. Sentences are divided into chunks and tagged by an external analyzer. The system then identifies the clauses, analyzes their relations and assigns them a tag indicating their syntactic nature. The system is characterized by its context-free grammar, proposing a set of tags for the "qu-" words adapted to our analysis, and by the use of the DCG formalism and the PROLOG language.

Keywords: partial syntactic analysis, syntactic clause, subordination, Prolog, CFG, DCG.

1. Introduction

Dans le cadre de travaux sur l'alignement de textes parallèles français-japonais, nous nous sommes fixé comme objectif l'alignement au niveau des propositions, afin de pouvoir fournir des données intéressantes pour diverses applications telles que la traduction automatique ou la linguistique contrastive. Face à la difficulté d'appariement due aux différences structurelles considérables du français et du japonais, nous avons posé comme hypothèse que les informations sur les relations entre les propositions seraient très utiles pour l'alignement de ces unités. Nous avons donc conçu le système SIGLé, réalisant non seulement la détection des propositions françaises avec indication de leur nature syntaxique, mais aussi l'analyse de leurs relations. Le système est caractérisé par sa grammaire de type CFG proposant un ensemble d'étiquettes adaptées à notre analyse pour les mots dits en « qu- ». La grammaire est écrite selon le formalisme DCG afin d'être intégrée directement dans le système implémenté en Prolog.

Le présent article est constitué d'un état de l'art, de l'exposé de notre grammaire et de notre système, et se termine par la présentation d'une évaluation.

2. Travaux antérieurs

Deux approches sont possibles pour reconnaître les propositions : soit en réalisant une analyse syntaxique complète de la phrase entrée, soit en se contentant d'une analyse morphologique et d'une analyse syntaxique partielle, telle qu'un *chunking*. Étant donné la difficulté extrême de création d'une grammaire complète, les systèmes de la détection automatique adoptent généralement l'approche reposant sur une analyse morphologique suivie d'une analyse syntaxique partielle.

Les méthodes proposées de détection des propositions – essentiellement pour l'anglais – sont classées en deux types. Le premier type recourt à des apprentissages automatiques comme la méthode proposé par (Ejerhed, 1988), ou les systèmes ayant participé à un *Workshop* de l'*ACL 2001* (Tjong *et al.*, 2001). Le deuxième type regroupe les méthodes avec grammaire basées sur le repérage des connecteurs (Abney, 1990) (Papageorgiou, 1997) (Leffa, 1998). Par ailleurs, la détection des propositions peut être une sous-tâche de l'opération principale et intégrée dans un autre cadre, tel que les travaux sur l'analyse rhétorique (Marcu, 1997) dans laquelle les propositions sont considérées comme unités élémentaires, et détectées également à l'aide d'une expression régulière basée sur le repérage des marqueurs discursifs.

Les travaux sur le français sont quasi absents et nous ne pouvons citer que (Maegaard et Spang-Hanssen, 1973) basés sur une grammaire du type « *transition network grammar* ». En revanche, il existe plusieurs méthodes d'analyse syntaxique partielle qui, sans être spécifiquement destinées à cette tâche, pourraient aider à repérer les propositions. Toutefois, selon la nature de leur partialité, ce type de système ne permet pas forcément de reconnaître les propositions. Par exemple, le système Syntex (Bourigault *et al.*, 2005)[1] fournit des analyses très détaillées au niveau des constituants plus petits que la proposition, mais très partielles au niveau de l'ensemble d'une phrase : le résultat d'une phrase correspond souvent non pas à un arbre mais à une forêt dont certains arbres sont constitués d'un seul élément.

Les méthodes proposées pour la détection des propositions fournissent déjà des résultats intéressants. Toutefois, sans parler de la langue traitée, ces approches consistant en une détection des frontières, ne permettent pas une analyse structurale déterminant les relations syntaxiques entre les propositions. De plus, du fait en particulier du caractère ambigu des connecteurs, l'expression régulière ne permet pas toujours de reconnaître correctement ces frontières et ce en particulier dans le cas de structures imbriquées. Nous avons donc conçu un système destiné spécifiquement à la détection des propositions du français, réalisant également l'analyse des relations entre les propositions. À cet effet précis, nous avons choisi l'utilisation d'une CFG contrairement aux travaux existants utilisant des expressions régulières. L'implémentation est réalisée en Prolog permettant ainsi, à partir d'une CFG écrite en formalisme DCG, de créer facilement un *parser*, sans concevoir aucun moteur particulier pour l'analyse syntaxique.

3. Grammaire pour la détection des propositions

Afin de construire les règles permettant la reconnaissance des propositions de manière économique à partir des résultats de *chunking*, nous avons défini trois éléments primaires : syntagme verbal (sv), syntagme nominal (sn) et autres compléments (cmp). Notre grammaire est constituée de plus de 160 règles avec une cinquantaine de non-terminaux, dont les trois éléments primaires. Elle repose sur deux axes d'études linguistiques : la typologie des propositions selon leur posi-

[1] Nous remercions D. Bourigault qui a eu la gentillesse de nous fournir les résultats d'analyse de nos corpus.

tion et la typologie des connecteurs. En effet, à peine avons-nous commencé à l'écrire, nous nous sommes rendu compte que les étiquettes « classiques » attribuées aux connecteurs de subordonnées n'étaient pas forcément adaptées à nos traitements. Nous avons alors étudié les travaux linguistiques de (Le Goffic, 1993) sur les connecteurs « qu- » pour aboutir à une typologie des subordonnées et des connecteurs adéquate pour notre grammaire de détection des propositions.

Mais, avant d'aborder ces études, il nous faut clarifier la notion de proposition, ce terme recouvrant en effet différents types d'unités, selon le point de vue de la personne qui l'utilise.

3.1. Définition des propositions

En nous appuyant sur la définition de (Le Goffic, 1993), nous désignons par proposition toute unité purement syntaxique, éventuellement non autonome, à structure phrastique constituée d'un sujet et d'un prédicat. Nous en définissons quatre types : 1. **principale**, 2. **subordonnée** introduite par un connecteur de subordination, 3. **coordonnée** introduite par un connecteur de coordination, 4. **détachée** entourée des deux séparateurs. Les subordonnées se distinguent elles-mêmes en plusieurs types comme présenté dans la section 3.2. Les propositions ainsi définies sont repérées principalement par la présence d'un verbe fini et de connecteurs.

Par ailleurs, de nos jours, l'existence d'éléments extérieurs à l'opposition sujet-prédicat dans une phrase est bien connue. Ce sont par exemple les thèmes, les introducteurs de cadres (Charolles, 1997) ou encore les constructions détachées représentant la prédication seconde (Combettes, 1998). Ces éléments extra-prédicatifs, souvent détachés, apparaissent notamment en début de phrase. Nous avons donc extrait également ces syntagmes à structure non-phrastique détachés en tête[2]. Nous supposons que ce choix s'avère particulièrement bénéfique lors de l'alignement avec les textes en japonais.

3.2. Typologie des subordonnées selon leur position dans la phrase

Par l'examen précis de la nature des connecteurs qui enchâssent les subordonnées dans la phrase, Le Goffic distingue quatre types de subordonnées avec connecteur en « qu- » : percontative, intégrative, relative et complétive. Il présente la correspondance de ces quatre notions avec les dénominations usuellement utilisées comme suit : 1. **percontative** (interrogative/exclamative indirecte) ; 2. **intégrative** (relative sans antécédent, circonstancielle en « qu- » ou « si ») ; 3. **relative** (relative avec antécédent) ; 4. **complétive** (complétive).

Mais, il est également possible de les classer selon leur position dans la phrase. En effet, les propositions subordonnées, fonctionnellement équivalentes à des termes simples, assument différentes fonctions dans la phrase (sauf celle de verbe) et leur position a un rapport avec la fonction qu'elles y jouent.

Ainsi, comme base de notre grammaire, nous définissons la segmentation suivante des subordonnées en quatre types selon leur position – donc la fonction qu'elles jouent – dans la phrase (le type de subordonnée selon Le Goffic est marqué entre parenthèses) :

– position pré-verbale : **proposition de sujet**
  qui dort dîne (intégrative)
  que vous avez menti me déçoit (complétive)

[2] Quoique conscients de leur existence, nous n'extrayons pas pour le moment les éléments extra-prédicatifs (à structure non-phrastique) apparaissant à une autre position que la position initiale : leur extériorité est parfois beaucoup moins nette et surtout la détection est plus délicate, nécessitant davantage d'information.

qui a peint ce tableau n'a jamais été découvert (percontative)
– position post-verbale : **proposition de complément** (subQ ci-après)
je pense qu'il viendra (complétive)
je me demande pourquoi il n'est pas venu (percontative)
je me moque de comment il a réussi (percontative)
– positions initiale et finale : **proposition circonstancielle** (subP ci-après)
quand je suis arrivé, il était déjà rentré (intégrative)
si tu ne manges pas, tu ne guériras pas (intégrative)
que le gouvernement propose une nouvelle loi, l'opposition crie au scandale. (complétive)
j'étais si fatigué que je me suis endormi très vite (intégrative)
– position post-nominale : **proposition déterminante** (subR ci-après)
la peinture qui m'a fascinée (relative)
l'idée que tout est fini (complétive)
son incertitude s'il devait obéir (percontative, tirée de (Le Goffic, 1993))
la déception du père quand il a entendu cette nouvelle (intégrative)

3.3. Typologie des connecteurs

En nous basant sur l'étude des positions précédentes, nous avons réalisé une classification des connecteurs, mots en « qu- », dont la synthèse est présentée dans le tableau 1.

	Positions initiale/finale (intégrative)	Position pré-verbale (intégrative, complétive percontative)	Position post-verbale (complétive, percontative)	Position post-nominale (relative, complétive percontative, intégrative)
qui		√	√	√
que	√	√	√	√
dont				√
où	√		√	√
quand	√		√	√
comme	√		√	√
si	√		√	√
quoi			√	√
lequel			√	√
quel			√	
combien			√	
comment			√	
pourquoi			√	

Tableau 1. Connecteurs du français

De cette classification, nous avons défini les quatre types de connecteurs suivants :
– **connecteurs isolés** (comportement particulier) : qui, que, dont ;
– **connecteurs ambigus** (apparaissant à 3 positions sauf pré-verbale) : où, quand, comme, si ;
– **indicateurs de propositions** (apparaissant seulement en position post-verbale) : quel (et ses formes fléchies), combien, comment, pourquoi ;
– **connecteurs relatifs** (apparaissant en position post-verbale et en position post-nominale) : quoi, lequel (et ses formes fléchies).
Nous avons alors pu définir une série d'étiquettes pour ces connecteurs, parfaitement adaptées

à un traitement de détection des propositions : *[qui]*, *[que]*, *[dont]* pour les connecteurs isolés, *[camb]* pour les connecteurs ambigus, *[ip]* pour les indicateurs de propositions et *[rel]* pour les connecteurs relatifs. Par ailleurs, les locutions conjonctives telles que « avant que », « bien que », « parce que » sont étiquetées *[cs]* (connecteur de subordination) et définies comme introduisant une proposition circonstancielle apparaissant en position initiale/finale. De plus, il existe encore un autre type qui introduit, non pas les subordonnées, mais les coordonnées, *[cc]* (connecteur de coordination) regroupant les mots dits traditionnellement conjonctions de coordination.

4. Réalisation

4.1. Fonctionnement de SIGLé

Notre système de détection des propositions SIGLé reçoit comme données les résultats du *tagging* et du *chunking* du texte entré réalisés par deux moyens extérieurs, un *tagger* et un *chunker* conçus tous les deux à l'Université Paris 7. SIGLé lui-même est constitué d'un module principal, de trois petits modules de pré-traitement et enfin d'un module de post-traitemant.

Les deux premiers modules de pré-traitement interviennent sur les données en entrée, reçues comme résultats des moyens extérieurs. Ils corrigent des erreurs évidentes détectables seulement avec les informations du contexte immédiat, et modifient certaines étiquettes en particulier celles des mots en « qu- ». L'utilisation de ces étiquettes permet, sans trop compliquer la grammaire, d'examiner toujours les deux possibilités syntaxiques que possèdent ces connecteurs : introductions d'un syntagme et d'une proposition. Ce choix est d'autant plus intéressant qu'il nous libère du risque de blocage dû à un étiquetage erroné de ces mots, certains très polysémiques, difficiles à réaliser sans une analyse syntaxique plus large que celle avec le simple contexte immédiat. Le troisième module de pré-traitement sert à mettre sous forme de liste Prolog les données entrées, et le module de post-traitemant, les résultats Prolog sous forme XML comme présenté figure 1.

```
<s id='1'>
<prop id='1' etq='principale' pere='0' fils ='2;'>
L'opposition considère [subQ]
</prop>
<prop id='2' etq='subQ' pere='1' fils =">
que la proposition du gouvernement néglige la population marginale
</prop>
</s>
```

Figure 1. Résultat en XML

Le module principal réalise la détection de propositions en exécutant une grammaire CFG écrite sous la forme DCG afin de pouvoir être compilée comme code PROLOG. Le module principal ne fournit comme résultat que la première réponse obtenue, ce qui rend pertinent l'ordre des règles. Nous discuterons des conséquences engendrées par ce choix dans la section 5. Le résultat fourni représente la structure schématique des phrases en propositions avec indication de leur nature syntaxique.

4.2. Transformation en DCG

La possibilité qu'offre la DCG d'ajout d'arguments au non-terminaux comme les prédicats, permet d'obtenir comme résultat d'analyse un arbre indiquant non seulement des propositions re-

connues, mais aussi leurs relations. Par exemple, la règle définissant la proposition prop→sujet, predicat. est réécrite comme :

```
proposition([prop, SUJ, PRED]) --> sujet(SUJ), predicat(PRED).
```

Chaque prédicat de la partie gauche de la règle prend comme argument une liste représentant le sous-arbre reconnu par la règle. La constante située à la première position dans la liste est l'étiquette attribuée à ce sous-arbre reconnu. Dans l'exemple, le prédicat proposition a comme argument la liste représentant le sous-arbre correspondant à la proposition reconnue constituée d'un sujet et d'un prédicat, et l'étiquette de ce sous-arbre est le premier élément de la liste, prop. Une fois toutes les règles mises sous cette forme, l'analyse de la phrase :

> L'opposition$_{[np]}$ considère$_{[vfin]}$ que$_{[que]}$ la proposition du gouvernement$_{[np]}$ néglige$_{[vfin]}$ la population marginale$_{[np]}$.

nous donne en sortie du programme Prolog[3] :

```
[s, [sujet, [sn, np]], [predicat, vfin, [subQ, [connect, que], [prop, [sujet, [sn, np]], [pred, vfin, [sn, np]]]]]]
```

4.3. Implémentation en PROLOG

L'analyse par une grammaire en DCG selon le mécanisme de PROLOG possède deux grands inconvénients : le problème de récursivité à gauche et celui de répétition des mêmes calculs due aux retours en arrière. Toutefois, à ces deux problèmes ont déjà été proposées plusieurs solutions par différentes techniques de compilation. Par exemple, le système SAX (Matsumoto et Sugimura, 1986) transforme une grammaire écrite en DCG en un programme PROLOG d'analyse syntaxique basée sur l'algorithme *Bottom-up Chart Parsing*, ce qui permet non seulement de résoudre le problème de récursivité à gauche mais aussi d'éviter la répétition des mêmes calculs.

Étant donné la taille relativement restreinte de notre grammaire, nous avons choisi une solution plus simple. Le problème de récursivité à gauche est résolu par inspiration de la forme normale Greibach. Il est déjà connu qu'elle permet d'éviter le problème de récursivité à gauche. Afin d'éviter les calculs redondants dus aux retours en arrière, nous avons adopté la stratégie de l'analyse tabulaire descendante en développant un interpréteur basé sur la méthode proposée dans (Pereira et Shieber, 1987). Cet algorithme permet de réutiliser des résultats de calculs déjà réalisés, par consultation des résultats stockés avant de commencer un nouveau calcul, qui s'avère identique.

5. Évaluation du système

Une première évaluation a été réalisée avec quatre corpus : un texte du sommet G8, *How to Unicode*, un extrait d'une œuvre littéraire (*Zadig* de Voltaire) et un corpus journalistique constitué de 15 articles du Monde Diplomatique. Le tableau 2 présente le résultat de l'évaluation[4].

[3] Dans l'analyse réelle, interviennent plus de prédicats n'ayant pas d'influence directe sur la détection des propositions, que nous avons supprimés de la liste de résultat pour favoriser la lisibilité.

[4] Le rappel est défini comme la proportion du nombre de phrases dont l'analyse a abouti sur le nombre total de phrases. La précision 1 est définie comme la proportion du nombre de phrases dont les frontières de propositions sont correctement détectées, sur le nombre total d'analyses de phrases ayant abouti. La précision 2 correspond à la proportion du nombre de phrases dont les relations des propositions sont correctement analysées, sur le nombre total de phrases dont les frontières sont correctement détectées. La précision totale est calculée par prec.1 $\times$ préc. 2, et correspond donc au taux de phrases dont la détection des frontières et l'analyse des relations sont correctes sur le nombre total d'analyses de phrases ayant abouti.

	Nb de phr	Rappel	Préc. 1	Préc. 2	Préc. totale
G8	53	96,2	98,0	100,0	98,0
Unicode	274	85,8	96,2	97,8	94,1
Zadig	1206	85,8	92,8	95,3	88,4
LMD	1713	84,9	89,2	98,0	87,4

Tableau 2. Résultat de la détection des propositions

Les taux de rappel sont relativement bas, mais la grande majorité des échecs provient du résultat erroné des pré-traitements (segmentation et *tagging*) à savoir plus de 90% pour G8 et Unicode, ce qui nécessitera un travail supplémentaire d'amélioration des modules de pré-traitement. D'autres erreurs sont dues à l'absence de règles adéquates.

L'évaluation de la précision est réalisée en deux temps : du point de vue de l'analyse linéaire (Préc.1 dans le tableau) concernant juste la détection des frontières de propositions ; et du point de vue de l'analyse structurale (Préc.2) déterminant les relations entre les propositions. L'utilisation d'une CFG a permis l'analyse correcte des structures imbriquées, difficiles à résoudre pour les méthodes avec une expression régulière. La phrase à propositions multiples :

> *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*

a été analysée sans problème comme suit :

(1) `principale :` [2] , *la triste vérité est* [3]
(2) `subP :` *si ces chiffres peuvent susciter l'étonnement*
(3) `subQ :` *que les habitants de Reay Road et des autres poches de misère* [4] *n'ont pas mieux où aller*
(4) `subR :` *qui prolifèrent*

Dans l'**analyse linéaire**, les erreurs sur les détections de frontières (marquées par « | ») non dues aux traitements antérieures se limitent essentiellement aux phrases contenant plusieurs virgules, notamment dans les structures de coordination. Par ailleurs, dans certains cas, bien qu'assez limités, un connecteur introduisant une proposition est interprété comme précédant un syntagme (ou vice versa), perturbant alors l'ensemble de l'analyse de la phrase.

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

Dans cette phrase, « que » a été interprété comme introducteur d'un syntagme et non d'une proposition. En effet, la règle définissant la phrase constituée d'un sujet et d'un prédicat est prioritaire sur les autres types de phrases dans l'ordre d'application.

Pour l'**analyse structurale**, nous constatons trois types d'erreurs.

1. Coordination des suboordonnées mal analysée faute de relatif
Le système est incapable de traiter, en tant que telles, les suboordonnées coordonnées sans pronom. Dans la phrase suivante, la proposition (3) est analysée comme coordonnée avec la proposition (1) alors qu'elle est coordonnée avec (2).

> *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*$_{(3)}$

L'introduction de traits grammaticaux simples tel que la personne et le nombre permettrait une analyse correcte, mais il existe également beaucoup de cas où elle ne suffit pas comme dans la phrase d'exemple précédente.

2. Coordination des subordonnées mal analysée du fait de l'ambiguïté de la virgule
Dans la mesure où une suboordonnée non coordonnée peut tout à fait être précédée par une

virgule et que la coordination des suboordonnées peut également être réalisée par une virgule, la suboordonnée précédée par une virgule est ambiguë pour le système et l'analyse par la subordonnée simple non coordonnée est prioritaire, entraînant ainsi des résultats erronés. Par exemple, dans la phrase :

> *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*$_{(5)}$

la deuxième subordonnée (3) est correctement interprétée comme complétive coordonnée à la première (2), mais la troisième (4), faute de conjonction de coordination, est mal analysée : elle l'est comme étant une simple relative précédée par une virgule. Afin de résoudre ce problème, une analyse plus précise est nécessaire telle que celle permettant de distinguer les relatives des complétives ou de déterminer l'antécédent des relatives pour interpréter correctement leur coordination. Cependant, ce type de calcul est très coûteux et risque de rendre le système peu opérationnel.

3. Faux étiquetages de subordonnées
Ils représentent plus de la moitié des erreurs dans la totalité des corpus analysés. Les connecteurs pouvant introduire différents types de subordonnées, dans certains contextes, le système interprète mal le type de subordonnée, et parfois même la structure relationnelle. Par exemple, toutes les subordonnées circonstantielles introduites par un connecteur *Camb* sont traitées comme des relatives quand elles suivent un syntagme nominal ou prépositionnel (ils détestent le peuple américain quand il ne leur ressemble pas), et comme des subordonnées de complément quand elles suivent un verbe (c'est facile à dire quand on n'est pas concerné dans sa chair). La détermination du type de subordonnée est parfois très difficile et nécessite des informations sémantiques voire extra-linguistiques.

Il existe un cas particulier de **relations ambiguës**. Lorsque la phrase contient trois propositions (ou plus), le rattachement de la troisième peut être ambigu et il est difficile dans ce cas de réaliser une évaluation de résultat. Ainsi, nous avons considéré certaines phrases comme ambiguës et ne les avons pas comptées parmi les erreurs. Par exemple, dans la phrase :

> *Paris avait estimé, à l'époque* | *, qu'une référence aux valeurs religieuses n'était pas acceptable* | *car elle soulevait des problèmes politiques et constitutionnels en France.*

le système l'interprète dans ce cas comme coordonnée à la principale (Paris avait estimé X car Y), mais elle peut tout à fait être rattachée à la subordonnée (une référence aux valeurs religieuses n'était pas acceptable car Y).

Par ailleurs, **le temps de calcul** avec l'interpréteur en analyse tabulaire est incomparablement plus rapide. Mais l'utilisation de mémoire est déjà très importante, et si nous envisagions l'introduction de plus d'informations, serait impératif le recours à un autre algorithme plus efficace.

6. Conclusion et perspectives des travaux futurs

L'utilisation d'une grammaire CFG permet de développer un système réalisant non seulement la détection des propositions mais aussi l'analyse de leurs relations, qui fournit des résultats intéressant pour l'alignement des propositions. Quoique plus lourde à manipuler qu'une expression régulière, une CFG destinée spécifiquement à la détection des propositions pouvant

être restreinte, s'avère plus facile à concevoir et surtout à améliorer qu'une grammaire d'analyse complète, nous paraissant une piste intéressante à continuer à explorer. Notre grammaire possède encore plusieurs possibilités d'améliorations. La manipulation de plus d'informations grâce à l'utilisation d'une forme augmentée permettrait certainement des améliorations, mais il est également important de tracer la limite : l'introduction excessive d'informations entraîne une complication des calculs, ce qui risque d'annuler les avantages du système partiel.

Nous allons terminer cet article avec les perspectives pour l'alignement des propositions. Notre hypothèse est d'aligner les propositions d'une paire de phrases à l'aide d'un graphe de relations, construit une fois la détection des propositions terminée. Les arcs des graphes sont typés avec les étiquettes – attribuées aux propositions par SIGLé – indiquant leur caractère syntaxique, auxquelles sont rajoutées ensuite les informations sémantiques que nous pouvons obtenir avec leur connecteur. La figure 2 présente un prototype d'alignement d'une paire de phrases français/japonais réalisé à l'aide des graphes de relations des deux phrases.

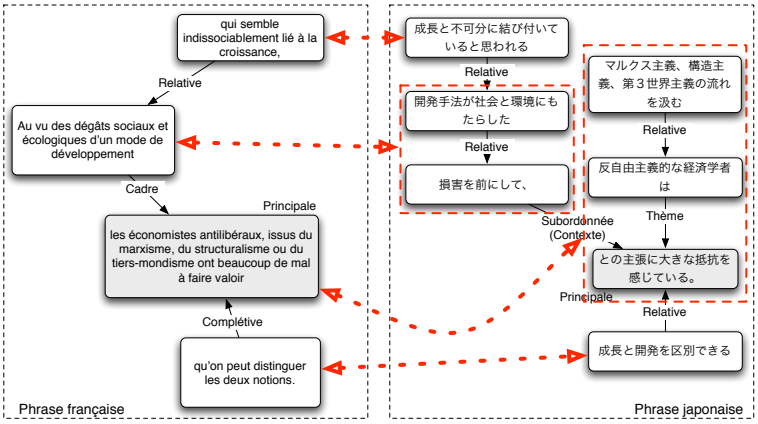

Figure 2. Alignement des propositions réalisé avec un graphe de relations

Références

ABNEY S. (1990). « Rapid Incremental Parsing with Repair ». In *Proceedings of the 6th New OED Conference*. University of Waterloo, Waterloo, Ontario.

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P. et OZDOWSKA S. (2005). « Syntex, analyseur syntaxique de corpus ». In *TALN 2005*.

CHAROLLES M. (1997). « L'encadrement du discours : univers, champs, domaines et espaces ». In *Cahier de recherche linguistique*, 6, 1 – 73.

COMBETTES B. (1998). *Les constructions détachées en français*. Ophrys, Paris.

EJERHED E. (1988). « Finding clauses in unrestricted text by finitary and stochastic methods ». In *ACL Proceedings, Second Conference on Applied Natural Language Processing*. p. 219–227.

LE GOFFIC P. (1993). *Grammaire de la phrase française*. Hachette, Paris.

LEFFA V. J. (1998). « Clause Processing in Complex Sentences ». In *Proceedings of LREC'98*.

MAEGAARD B. et SPANG-HANSSEN E. (1973). « Segmentation of french sentences ». In *Proceedings of COLING '73*.

MARCU D. (1997). « The rhetorical parsing of natural language texts ». In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97/EACL '97)*.

p. 96 – 103.

MATSUMOTO Y. et SUGIMURA R. (1986). « ronri-gata gengo ni motozuku kôbun kaiseki sisutemu SAX [Analyseur syntaxique basé sur un langage logique] ». In *Computer Software*, 3 (4), 308 – 315.

PAPAGEORGIOU H. (1997). « Clause recognition in the framework of alignment ». In R. Mitkov et N. Nicolov (éds.), *Recent advances in natural language processing*, p. 417 – 425. John Benjamins.

PEREIRA F. C. et SHIEBER S. M. (1987). *Prolog and Natural-Language Analysis*. CSLI, Stanford.

TJONG E. F., SANG K. et DÉJEAN H. (2001). « Introduction to the CoNLL-2001 Shared Task : Clause Identification ». In *Proceedings of CoNLL-2001*.