Groupes Nominaux Prédicatifs : utilisation d'une grammaire de liens pour l'extraction d'information

Jean Royauté, Élisabeth Godbert, Mohamed Madhi Malik

Université de la Méditerranée – LIF et CNRS
{royaute, godbert, malik}@lidil.univ-mrs.fr

Résumé

L'identification des structures prédicatives présente un grand intérêt quand on se situe dans une problématique d'extraction d'information. Si une littérature abondante existe à ce sujet, particulièrement dans le domaine de la génomique, la plupart des travaux portent sur les relations autour du verbe. Peu s'intéressent à la relation qui peut unir une nominalisation et ses actants dans un groupe nominal à tête prédicative (GNP). Nous montrons la complexité des différents types de GNP et des relations paraphrastiques qui les unissent avec les formes verbales, afin de donner une vue unifiée des structures prédicatives nomino-verbales. Nous montrons ensuite comment nous avons conçu une grammaire de liens permettant l'identification de chacun des actants dans les GNP. Nous en décrivons la mise en œuvre avec le Link Parser, pour l'extraction d'information dans des articles scientifiques du domaine de la Biologie.

Mots-clés : nominalisation, groupe nominal prédicatif, marqueurs prépositionnels, extraction d'information.

Abstract

The identification of predicative structures is of great interest in information extraction. Although there is abundant literature on this subject, particularly in the genomic field, the majority relates to the relations around the verb. Few are interested in the relation which can link a nominalization and its actants in a noun phrase with predicative head (NPP). Our work involves firstly showing the complexity of different types of NPPs and the paraphrastic relations which link them with the verbal forms, followed by a unified view of the nomino-verbal predicative structures. We further show how we designed a link grammar allowing the identification of each actant in the NNPs. We describe the implementation with Link Parser for information extraction in scientific articles in the field of Biology.

Keywords : nominalization, predicative noun phrase, prepositional markers, information extraction.

1. Introduction

La principale difficulté à laquelle est confrontée l'extraction d'information est de retrouver dans les textes des informations qui s'expriment en langue de manières très diverses. Dans ce cadre, nous nous intéressons ici à un phénomène syntaxique particulier, qui est actuellement peu ou pas traité dans les travaux sur l'extraction d'information : celui des groupes nominaux à tête prédicative (ou GNP).

La plupart des GNP sont formés à partir de nominalisations de verbes ou d'adjectifs. Suivant Pasero *et al.* (2004) nous disons qu'un nom est prédicatif quand il définit les mêmes relations argumentales qu'un verbe. Chacun des arguments joue un rôle conceptuel précis de type sujet, complément ou circonstant. Un GNP, par exemple *l'admiration de Louis pour Eve*, relié à la phrase *Louis admire Eve,* est ici formé d'une tête prédicative *admiration* suivie de ses actants

Louis et *Eve* précédés d'une préposition. On remarque qu'entre les deux structures, il y a conservation des actants et que l'on pourrait y adjoindre un éventuel circonstant (*depuis son enfance*). Elles diffèrent cependant par l'ordre d'apparition de la tête prédicative, de ses actants, de ses éventuels circonstants et par des marqueurs appropriés (prépositions ou conjonctions) qui permettent de localiser de façon stable les actants d'un GNP saturé.

Les variations syntaxiques nomino-verbales sont très fréquentes, tout particulièrement dans les textes scientifiques. Les chercheurs qui sont à l'origine de cette littérature y ont souvent recours, quand il n'est pas nécessaire d'inscrire dans le temps la relation argumentale qui unit un verbe ou un nom et ses actants (dephosphorylation of Cdc6 by PP2A).

Notre objectif est ici de définir une méthode d'analyse robuste des attachements prépositionnels que l'on rencontre dans ce type de GN, pour être en mesure d'en donner une représentation sémantique. Nous situons notre travail dans une perspective d'extraction d'information. Ce qui nous intéresse particulièrement, ce sont les relations que les mots entretiennent entre eux dans la phrase, plutôt qu'un découpage en syntagmes. Nous avons donc opté pour une analyse avec une grammaire de liens, qui est une variante des grammaires de dépendances (Tesnière, 1959) (Mel'cuk, 1988) : l'utilisation de ce type de grammaire est intéressante car elle permet de n'extraire que les schémas relationnels réellement utiles, et d'ignorer les autres. Nous utilisons l'analyseur Link Parser (Sleator et Temperley, 1991 ; www.link.cs.cmu.edu/link/) en y modifiant la grammaire de l'anglais que les auteurs ont développée.

Le domaine d'application auquel nous nous intéressons est l'analyse d'articles scientifiques, et en particulier de résumés ou d'articles de Biologie, rédigés en anglais. La demande des chercheurs en Biologie est très forte pour l'extraction automatique d'information et l'alimentation de bases de données ou bases de connaissances. Nous travaillons en particulier sur un corpus de résumés de MedLine qui portent sur des interactions protéines/gènes. Le traitement des GNP y est intéressant, car ces textes en sont riches : une évaluation rapide sur ce corpus nous a montré qu'environ 45 % des phrases décrivant des interactions contenaient des nominalisations. Ne pas faire un traitement pertinent des GNP nous ferait donc perdre une importante quantité d'information. Dans ce domaine, les noms prédicatifs les plus fréquemment utilisés sont *regulation, interaction, association*, *inhibition*, etc. Par exemple, on peut entre autres trouver les variantes suivantes autour de *regulation* : *regulation of X by Y, regulation of X, Y regulation by X ;* et l'on trouve aussi des GNP formés avec *up-regulation, down-regulation,* ou *co-regulation*.

Cet article est organisé comme suit. Dans la section 2 nous nous intéressons à une typologie des GNP significatifs, de leurs différents patrons et des formes verbales qui leur sont associées. Dans la section suivante nous donnons une vue unifiée des structures prédicatives tant du point de vue de la syntaxe que d'une représentation sémantique sous-spécifiée. Enfin, la section 4 est consacrée à la mise en place d'une grammaire de liens pour les GNP, son évaluation et la production de la structure informationnelle résultant de l'analyse.

2. Groupes nominaux prédicatifs

L'identification des relations prédicatives présente un grand intérêt quand on se situe dans une problématique d'extraction d'information. Le besoin qu'ont les biologistes de fouiller la littérature scientifique pour rechercher des schémas d'interaction génique destinés à alimenter des bases de données a réactivé cette problématique. Une littérature importante existe à ce sujet. Les principaux travaux, qu'ils portent sur des traitements réalisant une analyse complète (McDonald *et al*., 2004 ; Yakushiji *et al*., 2001) ou partielle de type shallow-parsing

(Alphonse *et al.*, 2004 ; Leroy et Chen, 2002) ou encore de type pattern-matching (Huang, 2004), reposent tous sur les structures verbales. Alphonse *et al.* (2004) s'intéressent aux structures prédicatives nomino-verbales, mais d'un point de vue général, sans décrire avec précision les différents patterns nominaux représentatifs de la complexité du problème. Concernant les nominalisations, bien que peu de travaux exploitent ces données, on retiendra le projet NOMLEX (Macleod *et al*., 1998), qui décrit finement environ 1000 nominalisations et leurs relations argumentales et qui a pu être utilisé dans des expérimentations d'extraction d'information (Meyers *et al.*, 1998). Nous situons notre travail dans cette perspective dans la mesure où il est fortement motivé linguistiquement et repose sur des données complexes.

Nous nous intéressons ici aux noms prédicatifs dérivés de verbes, et ignorons pour le moment les autres. Chacun de ces noms prédicatifs peut apparaître dans différentes formes de surface. Nous y distinguons d'une part les actants de la forme verbale associée, composés du sujet et des compléments essentiels, et d'autre part les circonstants. Nous montrons que la structure des GNP est étroitement corrélée à la nature du verbe qui correspond à la tête prédicative nominale. Nous reprenons la méthodologie utilisée dans Royauté (1999) et Pasero, Royauté et Sabatier (2004) sur les propriétés des GNP du français et donnons un premier inventaire des GNP de l'anglais, tels qu'on les rencontre dans la littérature en génomique.

Nous appuyons notre description des GNP de l'anglais, d'une part sur des observations en corpus (web, articles scientifiques, etc.) et d'autre part sur l'exploitation d'un lexique-grammaire (Specialist Lexicon, Browne *et al*. (2000) ; www.nlm.nih.gov/pubs/factsheets/umlslex.html) de l'anglais qui décrit les différents emplois verbaux (transitifs, intransitifs, prépositionnels, ditransitifs, infinitifs, à complétives, etc.) et qui donne pour chaque verbe la nominalisation à laquelle il est associé, ainsi que les prépositions pouvant être des introducteurs de compléments de nom. Leroy *et al.* (2002) se sont intéressés à de telles structures en utilisant des analyses locales et les données du Specialist Lexicon. Cependant les patrons qu'ils utilisent reposent essentiellement sur les verbes transitifs, mis ou non au passif et les constructions nominales associées. Nous utilisons ce lexique (que nous nommerons SL) dans une perspective différente. Partant du constat que SL, pour les nominalisations, ne nous donne aucune information de marquage des actants par les prépositions, il nous a été nécessaire de croiser les informations verbales avec les informations nominales pour restituer cette information de marquage qui nous fait défaut. Pour réaliser cet objectif, nous avons créé une base de données à partir de SL. Nous avons ensuite défini des classes verbales. Pour chaque type de verbes, nous choisissons un représentant verbe/nominalisation, dont nous cherchons en corpus les différentes formes nominales et verbales à partir des différents champs de SL. Nous limitons ces formes aux seuls actants. Ensuite nous sélectionnons tous les couples verbe/nom qui ont la même description.

Nous présentons ici six structures verbales, parmi les plus significatives, qu'il est possible de relier à des constructions nominales. Nous ne traitons pour le moment ni les verbes à complétives ni les verbes à infinitives, dont le type de complément peut apparaître également dans les constructions nominales. Ces structures représentent un sous ensemble de SL et des nominalisations de l'anglais. Dans les exemples qui suivent, nous adoptons les notations de M. Gross : N sans indice désigne un nom, N_0 N_1 … N_n désignent chacun un groupe nominal ayant une fonction syntaxique de sujet (par convention toujours d'indice 0) ou de complément.

- Classe de type N_0 V

Il s'agit des verbes strictement intransitifs. À ces verbes à un seul actant correspondent des nominalisations pour lesquelles seule la préposition *of* peut introduire le sujet :

N^{pred} of N_0 (*necrosis of the femoral head*)

Le représentant de cette classe est le couple : *necrose* / *necrosis.*

- Classe de type N_0 V N_1

Cette classe regroupe tous les verbes qui se construisent avec un COD et qui acceptent le passif. Certains de ces verbes ont également un emploi intransitif, cependant nous ne nous intéresserons qu'à la forme verbale saturée, car le COD même s'il est omis est virtuellement présent. Plus de 1000 couples verbe/nom entrent dans cette configuration. La forme nominale associée à cette construction est la suivante :

N^{pred} of N_1 by N_0 (*activation of protein kinase C delta by IFN-gamma*)

On remarque donc que la préposition *of* marque le COD et la préposition *by*, que l'on retrouve dans les phrases passives, marque le sujet. Toutes les nominalisations de SL associées à ce schéma de phrases acceptent ces prépositions comme complément de nom, sans le marquage que nous proposons. Le représentant de cette classe est le couple : *activate* / *activation*

À côté de ces formes régulières, il en existe d'autres, où à la préposition *by* peuvent être associées d'autres prépositions pour introduire le sujet (absorption : *absorption of glucose in/into/by/ the bloodstream*).

- Classe de type N_0 V Prép N_1

Ces constructions prépositionnelles, comme les constructions à COD, peuvent avoir un emploi intransitif. Cependant, dans ce cas aussi, le complément effacé est virtuellement présent. Les constructions prépositionnelles présentent un intérêt tout particulier dans la mesure où la préposition associée au verbe pour introduire le complément se retrouve à l'identique dans les constructions nominales comme ci-dessous :

N^{pred} of N_0 Prép N_1 (*fluctuation of tryptophans in gramicidin*)

On peut remarquer que contrairement aux nominalisations de type transitif, la préposition *of* n'introduit pas le complément mais le sujet. Le représentant de cette classe est le couple : *fluctuate* / *fluctuation.*

- Classe de type N_0 V N_1 Prép N_2

Il s'agit d'une construction transitive admettant un second complément qui dans un grand nombre de cas est optionnel (représentant de cette classe : *attribute* / *attribution*). La structure nominale saturée de ce type de nominalisation est la suivante :

N^{pred} of N_1 Prép N_2 by N_0 (*attribution of a protein fragment to a sequence by* N_0)

- Classe de type N_0 V Prép N_1 Prép N_2

Dans cette classe, les deux compléments font partie des entrées lexicales du verbe. Comme pour les nominalisations issues de verbes à un complément prépositionnel, on retrouve ces prépositions dans le GNP. Le GNP saturé aura la forme suivante :

N^{pred} of N_0 Prép N_1 Prép N_2 (*decrease of temperature from 200 K to 70 K*)

Le représentant de cette classe est le couple : *decrease* / *decrease.*

- Classe liée à des verbes relationnels : N_a V with N_b

Il s'agit d'une classe particulière dans la mesure où les actants peuvent être permutés du fait qu'ils appartiennent à la même classe sémantique (celle des objets concrets ou abstraits qui peuvent entrer en relation). Pour cette raison, nous les avons notés N_a et N_b. Le représentant de cette classe est le couple : *interact / interaction*. Plusieurs emplois nominaux équivalents, que nous détaillons ci-dessous, existent pour ce type de GNP.

N^{pred} of N_a with N_b	*(interaction of genes with proteins*)
N^{pred} of / between N_a and N_b	(*interaction of / between genes and proteins*)
N^{pred} of / between N_{plur}	(*interaction between two genes*)

Dans le dernier de ces emplois, N_{plur} désigne un nom au pluriel. La forme plurielle signifie que le nom en question représente une classe et que la relation s'établit entre deux ou plusieurs éléments de cette classe.

Pour tous les patrons nominaux que nous avons décrits, l'actant introduit par la préposition *of* peut se retrouver en position modifieur à gauche du nom prédicatif. Les actants marqués par une autre préposition peuvent difficilement occuper cette place, bien que nous en ayons observé quelques cas en corpus. C'est la raison pour laquelle nous acceptons de les analyser comme tels.

3. Structures prédicatives

Nous désignons par structure prédicative une classe structurée de prédicats nominaux et verbaux où viennent s'agréger les éléments d'information que l'on cherche à mettre en évidence. Ces structures prédicatives rendent compte à la fois des différentes formes de surfaces susceptibles d'être rencontrées et d'une représentation sémantique sous-spécifiée de l'information. À chacune de ces classes nous associons : (i) un ensemble de patrons syntaxiques nominaux-verbaux dont la singularité permet de définir la classe ; (ii) une (ou, exceptionnellement, plusieurs) structure argumentale qui une fois instanciée donne une structure informationnelle ; (iii) la liste des couples verbe/nom prédicatif qui appartiennent à cette classe.

3.1. Patrons syntaxiques

D'une façon générale, les patrons syntaxiques d'une structure prédicative correspondent aux différentes formes de surface qui véhiculent la même information et dans lesquelles le verbe/nom peut apparaître avec ses compléments essentiels. Nous désignons par patron canonique un squelette grammatical décrivant les différents actants du verbe dans sa forme saturée. À un patron canonique donné, il est possible d'associer des transformations particulières pouvant affecter l'ordre des mots et leur éventuel effacement. Ces opérations permettent d'engendrer toutes les variantes paraphrastiques de surface associées à un prédicat. Les patrons canoniques peuvent s'enrichir de compléments non essentiels, ou circonstants.

Ces patrons peuvent contenir des marqueurs appropriés, qui sont soit des prépositions soit des conjonctions, et qui précèdent certains actants. Pour un nom prédicatif, on définit ainsi des n-uplets de prépositions/conjonctions, dont la fonction est de marquer de façon stable les actants des groupes nominaux prédicatifs saturés. En cas d'effacement cette capacité de marquage s'amoindrit dans la mesure où ces marqueurs n'ont pas toujours la capacité d'identifier seuls les actants, particulièrement quand ils sont précédés de la préposition *of*.

La section précédente (Groupes nominaux prédicatifs) donne un inventaire de patrons représentatifs. Par exemple, les formes de surface les plus significatives des patrons liés au

couple *regulate/regulation* sont : *phosphorylation regulates proteins, proteins are regulated by phosphorylation, regulation of proteins by phosphorylation, proteins regulation by phosphorylation, etc.*

3.2. Structure argumentale, structure informationnelle

La structure argumentale est composée de plusieurs champs : le champ PRÉDICAT, puis un ou plusieurs autres champs (par exemple, AGENT, OBJET, etc.) qui correspondent aux arguments essentiels (les actants), et enfin un ou plusieurs champs CIRCONSTANT. Ce type de champ est destiné à contenir les compléments circonstanciels qui décrivent les conditions particulières de la réalisation du prédicat et de ses arguments.

En principe, à un prédicat correspondent plusieurs patrons syntaxiques, mais une unique structure argumentale. Pour certaines classes particulières de prédicats (par exemple les prédicats relationnels) on définit une seconde structure argumentale .

Par exemple, pour le couple *regulate* / *regulation*, la structure argumentale sera la suivante : << PRÉDICAT, AGENT, OBJET, CIRCONSTANT >>, et pour la classe représentée par le couple *interact* / *interaction*, on définira les deux structures argumentales qui suivent (cf exemples en 4.2) : << PRÉDICAT, CO-AGENT, CO-AGENT, CIRCONSTANT >> et << PRÉDICAT,AGENT-PLUR, CIRCONSTANT >>.

Cette structure argumentale peut être vue comme une représentation sémantique sous-spécifiée car les champs reçoivent des classes sémantiques neutres, immédiatement dérivables de la syntaxe. Ainsi le sujet reçoit le nom de AGENT, le ou les compléments, le nom de OBJET. Pour les verbes/noms locatifs, nous utilisons l'étiquette LOCALISATION. Pour certains de ces verbes/noms où les prépositions marquent explicitement un déplacement, nous marquons l'origine avec l'étiquette SOURCE et l'arrivée avec l'étiquette DESTINATION.

Lors du processus d'analyse syntaxique d'un GNP, certains patrons syntaxiques, saturés ou non, sont reconnus. Il est possible alors de produire une instance de la structure argumentale associée au patron reconnu. Cette instance, appelée structure informationnelle, pourra être, plus ou moins directement, intégrée dans une base de données.

Ainsi la structure informationnelle produite lors de l'analyse du GNP *proteins regulation by phosphorylation at two distinct sites* est*:*

PRÉDICAT = regulate/regulation
AGENT = phosphorylation
OBJET = proteins
CIRCONSTANT = at two distinct sites

Dans un GNP non saturé, on constate l'effacement d'un, de plusieurs ou de tous les actants. Dans ce cas, lors de l'instanciation, les champs correspondants n'apparaissant pas seront marqués « vide » dans la structure informationnelle. Par exemple, le champ AGENT est noté vide lors de l'analyse de la phrase *The downregulation of the c-myc promoter [was restored in SAOS2 cells]*.

4. Grammaire de liens et groupes nominaux prédicatifs

Les grammaires de liens sont une variante des grammaires de dépendances, issues des travaux de L. Tesnière (1959) et plus récemment de ceux de I. Mel'cuk (1988). Les grammaires de dépendance établissent dans la phrase des relations binaires spécifiques entre les mots : chaque mot, sauf la racine (qui est le verbe principal de la phrase), dépend d'un autre (son gouverneur). Le résultat d'une analyse en dépendances se présente sous la forme d'un arbre dont les arcs orientés portent des étiquettes désignant des fonctions grammaticales.

4.1. Les grammaires de liens

Nous avons choisi d'analyser les textes à partir d'une grammaire de liens. En tant que variante des grammaires de dépendances, elle établit des relations entre des paires de mots. On n'y trouve pas explicitement la notion de gouverneur, mais les étiquettes nous permettent par la suite de reconstruire les dépendances qui nous seront utiles. Les mots, dans l'implémentation de ce type de grammaire avec le Link Parser (Sleator et Temperley, 1991), sont reliés par la jonction entre un lien `X+` (vers la droite) et un lien `X-` (vers la gauche) où `X` est une étiquette quelconque. Par exemple, la grammaire ci-dessous montre que dans une analyse les noms propres sont reliés à un autre mot par un lien `Ss+` (fonction sujet) ou un lien `Os-` (fonction objet direct), et que le verbe *admire* est relié par des liens `Ss-` et `Os+`. La jonction d'un `Ss+` et d'un `Ss-` permet d'établir un lien `Ss` entre le sujet et le verbe, et de même pour `Os` entre le verbe et son complément. L'analyse de *Louis admire Eve* est représentée par le graphe ci-dessous.

```
Grammaire :                              +------Ss----+----Os----+
Louis Eve       : Ss+ or Os-             |            |          |
admire          : Ss- &  Os+             Louis      admire     Eve
```

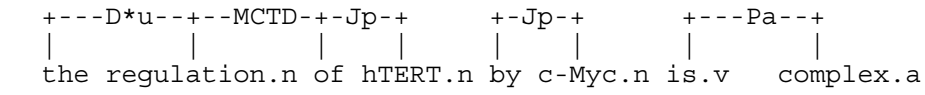
L'algorithme du Link Parser est en $O(n^3)$, ce qui correspond à la complexité des meilleurs algorithmes pour les grammaires hors-contexte. Cet algorithme impose d'une part que les relations s'établissent dans le demi-plan et qu'elles ne puissent pas se croiser (planarité) et d'autre part que tous les mots d'une phrase soient reliés entre eux (connectivité). Nous utilisons cet analyseur dont les sources et une grammaire conséquente de l'anglais sont disponibles. Chaque entrée du dictionnaire/grammaire est un couple (L,R), dans lequel L est une liste de mots et R est une formule plus ou moins complexe qui exprime l'ensemble des liens qui peuvent être attachés aux mots de L.

4.2. Définition de nouveaux liens pour les GNP

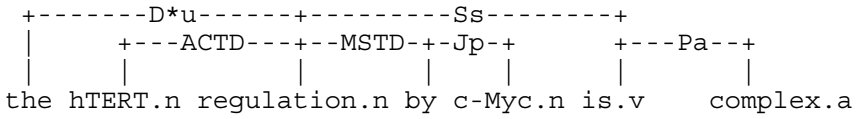
De façon standard, la grammaire du Link Parser permet le rattachement d'un verbe (resp. nom) à n'importe quelle préposition qui introduit un GN. Le lien utilisé est toujours MVp (resp. `Mp`) : modifieur verbal prépositionnel (resp. modifieur prépositionnel). Réciproquement, une préposition attend toujours un lien de type `Mp` ou `MVp`. Or, dans les GNP, des n-uplets de prépositions (couplés dans certains cas avec des conjonctions) précèdent et marquent les actants. Nous avons donc dû définir de nouveaux liens qui permettent d'identifier, lors de l'analyse d'un GNP, ses différents actants. Lors de l'analyse d'une phrase, ce sont ces liens qui sont recherchés, de façon préférentielle.

Ainsi, nous pouvons voir ci-dessous, dans le cas d'un GNP dont la nominalisation est liée à des constructions à COD (classe de type N_0 V N_1), que le lien `MSTD` identifie le sujet, tandis que le lien `MCTD` marque l'objet direct :

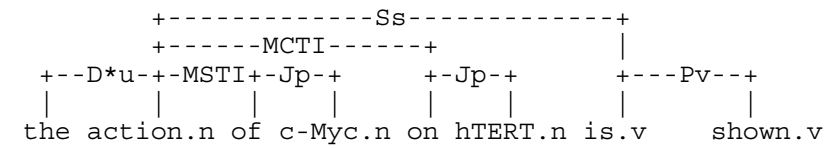
```
+--------------Ss--------------+
+-------MSTD-------+            |
```



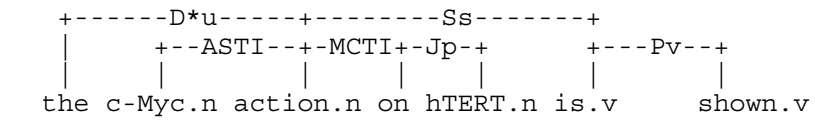
Un lien particulier (ACTD) a été créé pour traiter le cas où l'actant complément introduit par la préposition *of* se trouve en position de pré-modifieur. Remarquons qu'il existe à ce niveau une ambiguïté car exceptionnellement cette position peut être occupée par un circonstant.



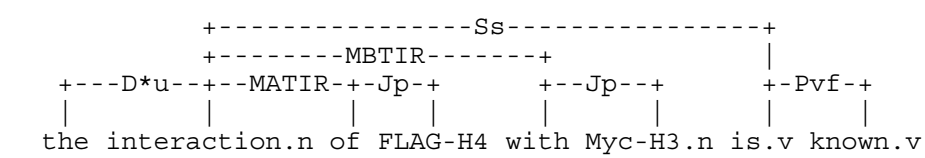
Concernant les nominalisations reliées à des verbes à complément prépositionnel (classe de type N_0 V Prép N_1), le lien MSTI identifie le sujet, tandis que le lien MCTI marque le complément :

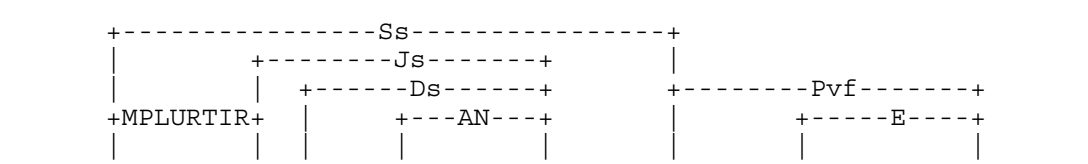


De la même façon que pour les constructions liées à des verbes à COD, le lien ASTI identifie l'actant introduit par la préposition *of* et placé en position modifieur. La relation identifiée cette fois-ci est une relation sujet.



Nous donnons ci-dessous deux exemples d'analyse avec des verbes relationnels (représentant : *interact* / *interaction*). Dans le premier, les liens MATIR et MBTIR identifient respectivement les co-agents de la classe N_a V with N_b , dans le second le lien MPLURTIR identifie l'agent pluriel de la même classe pour le patron N^{pred} of / between N_{plur} :





La définition des liens spécifiques pour l'analyse des GNP minimise les cas d'ambiguïté d'attachement prépositionnel. Néanmoins, il reste des ambiguïtés irréductibles, et nous avons intégré leur traitement dans la grammaire. Plutôt que produire plusieurs analyses différentes, nous avons choisi de signaler ces ambiguïtés par des étiquettes appropriées. Le cas le plus fréquent est celui où l'on trouve un nom en position de pré-modifieur, et aucun actant en

position de post-modifieur : l'étiquette `ASTDorACTD` exprime alors cette ambiguïté entre les fonctions sujet ou complément (*the hTERT regulation is ...*).

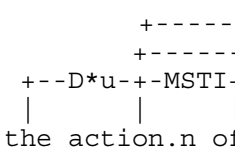
Parmi les ambiguïtés que nous ne sommes pas en mesure de traiter, signalons celle qui est attachée à la structure N_1^{pred} ... N_2^{pred} ... Prep... pour laquelle Prep figure dans les entrées lexicales de N_1^{pred} et N_2^{pred} (*the activation of the regulation of X by Y is...*). De même, dans le cas où un GNP se trouve dans la phrase en position complément, l'analyseur produit une double analyse : dans un premier temps, le groupe prépositionnel du GNP construit avec une préposition autre que *of* est analysé comme actant du GNP, et dans un second temps comme circonstant du verbe qui précède (*we show the absorption of glucose in the bloodstream)*.

Mise en œuvre et premiers résultats

Notre corpus est constitué de 1935 résumés de MedLine, annotés manuellement. Cette annotation nous a permis d'extraire 1337 phrases qui expriment une interaction entre deux gènes. Nous avons par ailleurs identifié 278 nominalisations verbales qui apparaissent dans ces phrases. L'intégration dans la grammaire de ces nominalisations a nécessité la création de 11 classes ou sous-classes, dont les plus significatives ont été présentées dans la section 2. Nous avons par ailleurs intégré dans le LP les entrées du Specialist Lexicon grâce aux données mises en ligne par P. Szolovits, MIT (Szolovits, 2003, www.medg.lcs.mit.edu/projects/text), et l'ensemble des noms de gènes et protéines de notre corpus. Pour les tests actuellement en cours, nous couvrons les différentes classes et sous-classes de notre grammaire avec les 58 nominalisations les plus fréquentes (*interaction, expression, transcription, activation, regulation*...). Ceci représente environ 1500 GNP à analyser. Nous avons procédé par sondage, en tirant de façon aléatoire 60 GNP dans lesquels la nominalisation apparaît avec au moins un actant. Nous retenons, parmi les analyses produites par le LP, la première qui contient le maximum de liens que nous avons créés spécifiquement pour les GNP. Les premiers résultats montrent que les rattachements prépositionnels se font correctement, dans la mesure où, sur notre échantillon, nous obtenons une précision de 88,5 %.

4.3. Production des structures informationnelles

À l'issue de l'analyse d'un GNP, les champs de la structure informationnelle sont remplis via les liens spécifiques aux GNP. Cette extraction se fait à partir de la forme linéaire que propose le Link Parser pour l'affichage du résultat de l'analyse. Par exemple, le résultat de l'analyse de *the action of c-Myc on hTERT is shown* peut être présenté sous les deux formes suivantes :




```
 (m)   the            D       <---D*u--->  D*u       action.n
 (m)   action.n       Ss      <---Ss---->  Ss        is.v
 (m)   action.n       MCTI    <---MCTI-->  MCTI      on
 (m)   action.n       MSTI    <---MSTI-->  MSTI      of
 (m)   of             J       <---Jp---->  Jp        c-Myc.n
 (m)   on             J       <---Jp---->  Jp        hTERT.n
 (m)   is.v           Pv      <---Pvf--->  Pvf       shown.v
```

La première forme permet de vérifier visuellement la cohérence globale de l'analyse. La seconde permet un traitement informatique qui, par un parcours sélectif des liens pertinents, construit la structure informationnelle suivante :

PRÉDICAT = action
AGENT = c-Myc
OBJET = hTERT
CIRCONSTANT = ϕ

5. Conclusion

Nous avons montré dans cet article la complexité des structures prédicatives nomino-verbales et l'intérêt d'en donner une description approfondie en extraction d'information, tant pour l'analyse syntaxique que pour capturer les éléments d'information pertinents à intégrer dans une base de connaissances. Cela nous a amenés à établir une typologie, reposant sur un sous ensemble significatif des GNP, avec leurs différents patrons et leurs structures argumentales. Pour tester la validité de notre travail, nous avons modifié la grammaire du Link Parser. Nous avons défini des liens spécifiques pour identifier le rôle de chaque actant dans la grammaire.

Des tests en cours, nous retirons la nécessité de compléter la description des structures prédicatives et de nous intéresser au problème des non attendus, afin de pouvoir réaliser des tests en vraie grandeur. Par exemple, pour le couple *interact*/*interaction*, les co-agents de la relation peuvent être mis en position modifieur, séparés par un slash ou un tiret (*the c-MYC-INI1 interaction was observed both in vitro and in vivo, [...] gp17/CD4 interaction*). Par ailleurs, un certain nombre de phénomènes linguistiques importants restent à traiter dans une perspective d'extraction d'information. Il s'agit de s'intéresser : (i) aux nominalisations d'adjectifs (*deficient/deficiency*) ; (ii) aux prédicats nominaux non dérivés (*analogy, identity*, etc.) ; et (iii) au traitement des formes pronominales ou possessives (*its action on hTERT*). Pour traiter ces différents cas, nous définirons des liens spécifiques similaires à ceux que nous avons présentés dans la section 4.

6. Remerciements

Nous sommes très reconnaissants à Christine Brun et Bernard Jacq, du Laboratoire de Génétique et Physiologie du Développement, Marseille, de nous avoir fourni un corpus annoté de résumés de MedLine sur lequel nous avons pu travailler.

Références

ALPHONSE E., AUBIN S., BESSIÈRES P., BISSON G., HAMON T., LAGUARIGUE S., NAZARENKO A., MANINE A-P., NÉDELLEC C., VETAH M.OULD ABDEL, POIBEAU T., WEISSENBACHER D. (2004). « Event-based information extraction for the biomedical domain: the Caderige project ». In *Proceedings of the International Workshop on Natural language, Processing in Biomedicine and its Applications (JNLPBA)* : 43-49.

BROWNE A.C., MCCRAY A.T., SRINIVASAN S. (2000). *The SPECIALIST lexicon technical report, Lister Hill National Center for Biomedical Communications*. National Library of Medicine.

HUANG M., ZHU X., HAO Y., PAYAN D.G., QU K., LI M. (2004). « Discovering patterns to extract protein-protein interactions from full texts ». In *Bioinformatics* 20 (18) : 3604-3612.

LEROY G., CHEN H., MARTINEZ J.D. (2003). « A shallow parser based on closed-class words to capture relations in biomedical text ». In *Journal of Biomedical Informatics* 36 : 145-58.

MACLEOD C., GRISHMAN R., MEYERS A., BARRETT L., REEVES R. (1998). « Nomlex: A lexicon of nominalizations ». In *Proceedings of the Eighth International Congress of the European Association for Lexicography*. Liège : 187-193.

MCDONALD D.M., CHEN H., SU H., MARSHALL B.B. (2004). « Extracting gene pathway relations using a hybrid grammar: the arizonarelation parser ». In *Bioinformatics* 20 (18) : 3370-3378.

MEL'CUK I.A. (1988). *Dependency Syntax : Theory and Practice*. State University of New-York Press.

MEYERS A., MACLEOD C., YANGARBER R., GRISHMAN R., BARRETT L., REEVES R. (1998). « Using NOMLEX to produce nominalization patterns for information extraction ». In *Proceedings of the COLING-ACL '98 Workshop on Computational Treatment of Nominals*. Montréal.

PASERO R., ROYAUTÉ J., SABATIER P. (2004). « Sur la syntaxe et la sémantique des groupes nominaux à tête prédicative ». In *Lingvisticae Investigationes* 27 (1) : 83-124.

ROYAUTÉ J. (1999). *Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information.* Thèse de doctorat, Université Henri Poincaré (Nancy 1).

SLEATOR D., TEMPERLEY D. (1991). *Parsing English with a Link Grammar. Carnegie Mellon University Computer Science technical report, CMU-CS-91-196*. Carnegie Mellon University.

SVOLOVITS P. (2003). « Adding a Medical Lexicon to an English Parser ». In *Proceedings of AMIA 2003 Annual Symposium* : 639-643.

TESNIÈRE L. (1959). *Eléments de syntaxe structurale*. Klincksieck, Paris.

YAKUSHIJI A., TATEISI Y., MIYAO Y., TSUJII J. (2001). « Event extraction from biomedical papers using a full parser ». In *Proceedings of the sixth Pacific Symposium on Biocomputing.*