TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Enrichissement de structures en dépendances par réécriture de graphes

Guillaume Bonfante, Bruno Guillaume, Mathieu Morey, Guy Perrier
INRIA Nancy-Grand Est - LORIA - Nancy-Université

Résumé. Nous montrons comment enrichir une annotation en dépendances syntaxiques au format du *French Treebank de Paris 7* en utilisant la réécriture de graphes, en vue du calcul de sa représentation sémantique. Le système de réécriture est composé de règles grammaticales et lexicales structurées en modules. Les règles lexicales utilisent une information de contrôle extraite du lexique des verbes français *Dicovalence*.

Abstract. We show how to enrich a syntactic dependency annotation of the *French Paris 7 Treebank* format, using graph rewriting, in order to compute its semantic representation. The rewriting system is composed of grammatical and lexical rules structured in modules. The lexical rules use a control information extracted from *Dicovalence*, a lexicon of French verbs.

Mots-clés : dépendance, French Treebank, réécriture de graphes, Dicovalence.

Keywords: dependency, French Treebank, graph rewriting, Dicovalence.

Introduction

Cet article propose une méthode d'enrichissement des structures en dépendances syntaxiques de surface et il applique cette méthode au *French Treebank de Paris 7* (par la suite noté FTB). Il entre dans la ligne de recherche ouverte par Bonfante *et al.* (2010) où nous montrions comment calculer — au moyen de la réécriture de graphes — la sémantique d'une phrase à partir de sa structure en dépendances syntaxiques. De manière plus générale, notre approche s'inscrit dans le contexte des méthodes exactes et symboliques de calcul en TAL.

Les représentations de la syntaxe en dépendances connaissent une popularité croissante pour l'évaluation et la comparaison d'analyses syntaxiques. Les raisons principales en sont données par Kahane (2001) : les dépendances syntaxiques sont lexicalisées et proches de la sémantique. Il existe très peu de corpus annotés en dépendances pour le français ; mais, récemment, Candito *et al.* (2009) ont montré comment produire une annotation en dépendances de surface du FTB à partir de son annotation en constituants (Abeillé *et al.*, 2003). Dans cet article, nous utilisons ce corpus pour tester notre système.

Dans Bonfante *et al.* (2010), nous avons proposé le principe de la réécriture de graphes pour calculer la sémantique à partir de la syntaxe. Nos entrées étaient des analyses syntaxiques profondes à la manière des structures tectogrammaticales du *Prague Dependency TreeBank* [1] (Hajič *et al.*, 2000). Dans notre cas, il s'agissait de structures enrichies du format PASSAGE [2]. Dans Bonfante *et al.* (2011), nous avons montré que l'on pouvait employer en fait le format FTB dès lors que certaines dépendances syntaxiques profondes étaient ajoutées : les arguments lexicalement ou grammaticalement déterminés des infinitifs et les antécédents des pronoms relatifs et réfléchis et

1. http://ufal.mff.cuni.cz/pdt2.0/
2. http://atoll.inria.fr/passage/

des sujets répétés. Dans les deux cas, les expérimentations demandaient une phase manuelle de préparation des données pour ajouter les annotations manquantes. Par ailleurs, nos systèmes surgénéraient car ils n'intégraient pas d'informations lexicales.

Nous pallions ici ces défauts en proposant un enrichissement automatique du FTB par des règles purement grammaticales et par d'autres qui utilisent de l'information lexicale extraite de Dicovalence (Van den Eynde & Mertens, 2003). Ces règles, en combinaison avec celles de l'article Bonfante *et al.* (2011), permettent donc d'*obtenir automatiquement une structure sémantique* [3] *à partir d'une annotation en dépendances provenant du* FTB.

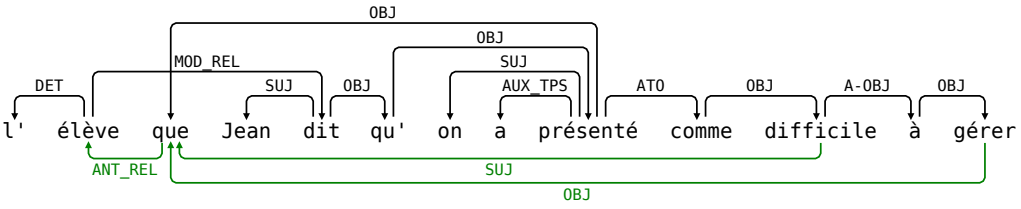
Les problèmes — échec de la réécriture, structures produites malformées — que nous avons rencontrés au cours de nos expérimentations, relèvent souvent d'erreurs d'annotations ou d'incohérences dans les structures du FTB. Notre système peut donc servir, par effet de bord, d'outil de fouilles d'erreurs et d'aide à la correction de corpus annotés.

Dans la section 1, nous posons le problème à travers un exemple introductif et nous présentons brièvement le modèle de réécriture de graphes choisi pour enrichir les annotations en dépendances du FTB. Puis dans la section 2, nous décrivons le système de règles de réécriture de graphes qui a été utilisé pour le faire et enfin dans la section 3 nous présentons la validation expérimentale qui a été effectuée sur le FTB.

1 Position du problème

De façon générale, pour calculer la sémantique, il est utile d'ajouter à l'annotation en syntaxe de surface deux types d'informations : les antécédents d'anaphores syntaxiques et certains actants syntaxiques. Dans la démarche proposée ici, la structure en dépendances de surface est celle du FTB. Nous nous imposons comme contrainte de ne pas modifier l'entrée mais seulement d'enrichir la structure à l'aide de nouvelles relations. Ce choix permet de rester complètement compatible avec d'autres travaux ou outils basés sur le format du FTB et de leur servir de guide. Il évite également de définir un nouveau format ad-hoc.

L'exemple donné ci-dessous est annoté par, en noir, les dépendances syntaxiques du guide d'annotation utilisé par (Candito *et al.*, 2009) [4] et, en vert, les dépendances à ajouter.



Pour les antécédents d'anaphores syntaxiques, nous nous en tenons à des phénomènes bien délimités : les antécédents des pronoms relatifs, des pronoms personnels réfléchis et des pronoms personnels répétitions de sujets. Retrouver l'antécédent d'un pronom relatif peut s'avérer complexe. Sur notre exemple, il faut remonter une chaîne de trois dépendances OBJ qui va du pronom relatif *« que »* jusqu'à la tête de la relative *« dit »* (en passant par *« présenté »* et *« qu' »*). De là, on retrouve l'antécédent *« élève »* en remontant la dépendance MOD_REL. Plus

3. une structure DMRS (Copestake, 2009)
4. http://www.linguist.univ-paris-diderot.fr/~mcandito/Rech/FTBDeps

généralement, la longueur des chaînes de dépendances n'est pas bornée. En outre, la chaîne peut contenir plusieurs sortes de relations et les structures de traits des nœuds rencontrés suivent certaines contraintes. Pour ces deux raisons, le calcul n'est pas immédiat.

Les actants syntaxiques à ajouter sont essentiellement les sujets des infinitifs s'ils sont présents dans la phrase. Nous mettons également dans cette catégorie les objets directs d'infinitifs comme dans la construction du *tough movement*. Ainsi dans l'expression *« un livre difficile à lire »*, l'objet de *« lire »* est *« livre »*. Manquent également dans le FTB les sujets des participes présents et des participes passés dans leur utilisation adjectivale. De façon plus générale, il paraît utile de traiter de façon uniforme les syntagmes adjectivaux et de leur associer systématiquement un sujet, que leur tête soit un participe ou un adjectif.

Comme les actants syntaxiques sont mutuellement dépendants, le calcul des relations SUJ et OBJ doit en tenir compte. Dans notre exemple, nous procédons dans l'ordre suivant. Nous commençons par marquer la dépendance SUJ entre l'adjectif *« difficile »* et son sujet *« que »*. En effet, *« que »* est l'objet direct de *« présenté »* et *« difficile »* en est l'attribut de l'objet. Partant de celle nouvelle relation, nous pouvons marquer la relation OBJ du verbe *« gérer »* vers le sujet de *« difficile »*. Cette annotation découle d'une propriété lexicale de l'adjectif *« difficile »* : son aptitude au *tough movement*.

Pour enrichir les structures de dépendances, nous employons le formalisme (β-calcul) que nous avons introduit dans Bonfante *et al.* (2010, 2011). Dans le cadre restreint de cet article, nous n'utilisons pas la réécriture de graphes de façon essentielle ; nos règle repèrent, dans des graphes, des motifs qui sont des arbres. Les seules transformations consistant à ajouter des arêtes, on pourrait certainement exprimer ces règles avec un formalisme moins puissant mais l'utilisation du β-calcul permet l'homogénéité de traitement avec les modules proposés dans les autres articles. Dans ce formalisme, un calcul procède par transformations successives des structures de dépendances jusqu'à leur normalisation. Le système est décrit par un ensemble fini de règles, chacune d'entre elles étant donnée par un motif et une liste de commandes élémentaires (ajout d'arête, suppression de nœud, etc). Une étape de calcul consiste à reconnaître le motif d'une règle dans le graphe courant et à modifier le graphe selon les commandes de la règle. La figure 1, en fin d'article, montre le déroulement de la réécriture sur la phrase *« Je trouve ce livre difficile à lire »*. L'encadré en haut de la figure présente les deux règles `ATTR-OBJ` et `TOUGH-MOVEMENT` utilisées dans cet exemple. Pour chaque règle, le schéma représente le *motif* reconnu, en rose et rouge. Au-dessous, est notée la liste de commandes effectuant la réécriture. Par exemple, pour la règle `ATTR-OBJ` , la liste se réduit à la commande *add_edge A –[SUJ]–> O* qui ajoute une relation SUJ de l'attribut de l'objet à l'objet du verbe.

Pour avoir un contrôle global sur le calcul, nous avons montré dans (Bonfante *et al.*, 2011) qu'une organisation modulaire des règles simplifie la tâche de développement d'un système de règles. De fait, des considérations linguistiques justifient à la fois la définition des modules et leur ordre d'application.

2 Règles utilisées pour l'enrichissement

L'enrichissement de l'annotation de surface se fait au moyen de quatre types de règles :
- des *règles grammaticales d'actants* déterminent de façon purement grammaticale certains sujets d'infinitifs, de participes et d'adjectifs,
- des *règles lexicales d'actants*, extraites d'un lexique, déterminent les sujets ou objets d'infinitifs compléments de verbes ou d'adjectifs à contrôle,
- des *règles d'antécédents* déterminent les antécédents des anaphores syntaxiques,
- des *règles de coordination* ajoutent les actants syntaxiques manquants aux conjoints.

2.1 Règles grammaticales d'actants

2.1.1 Sujets des syntagmes adjectivaux

Les participes têtes de syntagmes adjectivaux n'ont pas de sujets explicites dans le FTB car celui-ci ne contient que des arbres de dépendances. Le sujet d'un participe a déjà une autre fonction syntaxique ; le désigner reviendrait donc à lui attribuer un deuxième gouverneur. Voyons quelques exemples (les participes sont mis en gras) :

Jean arrive en **chantant**.
Abandonnée *de tous, elle ne sait plus que faire.*
Pierre sait Marie **délaissée** *par son mari.*
Un homme se **présentant** *comme son frère vient d'arriver.*

La configuration grammaticale dans laquelle se trouvent ces participes permet de déterminer leur sujet.

Nous étendons d'ailleurs la notion de sujet aux syntagmes adjectivaux qui ont un adjectif comme tête, pour deux raisons. Premièrement, au niveau sémantique, les adjectifs vont se traduire (comme les verbes) par des prédicats dont il faudra déterminer les arguments. Deuxièmement, les adjectifs peuvent se trouver dans des constructions variées : épithète, attribut du sujet, attribut de l'objet, dislocation. Toutes ces constructions sont compatibles avec le *tough movement*, comme le montrent les exemples suivants :

Je connais un livre **difficile** *à lire.*
Ce livre passe pour **difficile** *à lire.*
Je trouve ce livre **difficile** *à lire.*
Difficile *à lire, ce livre n'est pas à conseiller à tout le monde.*

Dans ces phrases, la relation entre *« livre »* et *« difficile »* implique que *« livre »* est objet de *« lire »*. Attribuer un sujet aux adjectifs permet de traiter les phénomènes comme le *tough movement* (Rezac, 2006) avec une seule règle, qui est indépendante de la configuration dans laquelle ces adjectifs apparaissent. Cela conduit d'ailleurs à une approche uniforme des adjectifs et des verbes, ce qui simplifie la définition des règles de calcul du sujet et des autres compléments.

Il y a autant de règles que de constructions intégrant des syntagmes adjectivaux. Ici, nous en considérons sept : épithète, attribut du sujet (avec ou sans préposition), attribut de l'objet (avec ou sans préposition), dislocation, gérondif. La règle `ATTR-OBJ` de la figure 1 (en fin d'article) s'applique aux attributs de l'objet sans préposition.

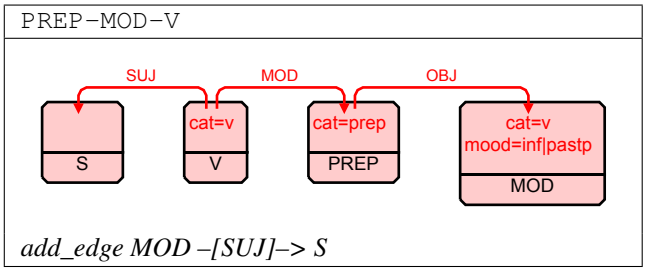
2.1.2 Sujets des infinitifs dans certaines constructions grammaticales

Comme les sujets des participes, les sujets des infinitifs ne sont pas exprimés dans le FTB ; les ajouter aboutirait également à violer la contrainte d'arbre des structures en dépendances. Certaines constructions syntaxiques associées à des mots grammaticaux contiennent pourtant des infinitifs dont le sujet n'est pas ambigu. Considérons les exemples suivants où les infinitifs concernés sont en gras :

Jean vient à Paris pour **travailler**.
Jean ne vient jamais à Paris sans **visiter** *la Tour Eiffel.*
Jean est trop poli pour ne pas **saluer** *Marie.*

Dans les deux premiers exemples, les prépositions *« pour »* et *« sans »* introduisent des compléments circonstanciels qui sont des infinitifs. Dans le dernier exemple, le couple *« trop . . . pour »* se construit avec un premier élément qui peut être un adjectif (ici *« poli »*), un adverbe ou un verbe et un second élément qui est toujours un infinitif (*« saluer »*).

Deux règles traitent ces différentes constructions. Dans le cas des infinitifs compléments circonstanciels, la règle `PREP-MOD-V` ci-contre fait apparaître que le sujet de l'infinitif est le sujet de la proposition principale modifiée [5]. Dans le cas de la construction *« trop + ADJ + pour + VINF »*, une autre règle fait apparaître le sujet de l'infinitif comme le sujet de l'adjectif. Là encore, attribuer un sujet aux adjectifs permet d'appliquer une même règle dans différentes configurations syntaxiques.



2.2 Règles lexicales d'actants

2.2.1 Le cas des verbes à contrôle ou à montée

Lorsqu'un infinitif est complément d'un verbe à contrôle ou à montée, son sujet, s'il est exprimé dans la phrase, est déterminé par ce verbe. Voyons quelques exemples où les infinitifs sont mis en gras et leur sujet profond est souligné.

Jean semble **changer** *d'avis.*
Jean permet à Marie de **venir**.
Jean promet à Marie de **venir**.
Jean propose à Marie de **venir**.

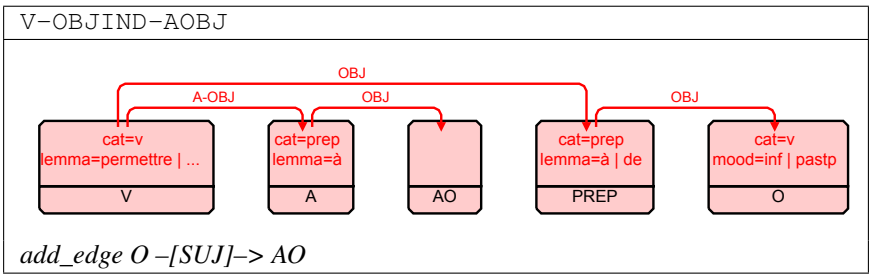
Dans la première phrase, *« sembler »* est un verbe à montée, l'infinitif qui le suit partage son sujet. Dans les deux suivantes, les verbes *« permettre »* et *« promettre »* déterminent le sujet de *« venir »* : *« Marie »* pour la première et *« Jean »* pour la deuxième. La dernière phrase est ambiguë, le sujet de *« venir »* peut être *« Jean »* ou *« Marie »*. Ce dernier exemple montre que le système que l'on a à décrire est nécessairement non confluent.

Les verbes à contrôle peuvent être groupés en différentes classes selon la fonction de l'infinitif contrôlé et selon la fonction du syntagme sujet de cet infinitif. Il y a trois possibilités pour chacun de ces deux facteurs : sujet, objet direct et objet indirect, donc six classes possibles (les deux fonctions syntaxiques sont nécessairement différentes). Pour chaque classe, il y a plusieurs règles, afin de gérer l'éventuelle présence d'un complémenteur *« à »* ou *« de »* introduisant l'infinitif.

Les verbes à contrôle possèdent, dans le lexique des verbes du français *Dicovalence* (Van den Eynde & Mertens, 2003), un ou plusieurs champs *PIVOT* qui contiennent leurs informations de contrôle. Ces informations sont extraites automatiquement de *Dicovalence* pour ancrer lexicalement les règles des verbes à contrôle, comme *« permettre »* dans l'exemple suivant :

```
VAL$    permettre: P0 P1 (P2)
NUM$    60700
FRAME$  subj:pron|n:[hum], obj:pron|n|compl|de_inf:[abs,mood:subj], ?objà:pron|n:[hum]
PIVOT$  P2/P0 [below de_inf in P1]
```

5. Il y a quelques exceptions à cette règle qui ne sont pas traitées : les constructions impersonnelles et certaines expression figées.



`P2/PO [below de_inf in P1]` se lit ainsi : l'objet indirect de *« permettre »* introduit par *« à »*, noté P2 dans *Dicovalence*, est le sujet, noté P0, d'une infinitive introduite par *« de »* (de_inf) qui est l'objet direct , noté P1, de *« permettre »*. Cela se traduit dans la règle `V-OBJIND-AOBJ` par la commande *add_edge O –[SUJ]–> AO* qui ajoute une relation SUJ de l'objet direct O vers le groupe nominal objet indirect AO.

2.2.2 Le cas des adjectifs à contrôle

Lorsqu'un infinitif est complément d'un adjectif à contrôle, son sujet ou son objet — quand il est présent dans la phrase — est déterminé lexicalement. Le *tough movement* est l'une de ces configurations. Voici quelques exemples où les infinitifs sont en gras :

Jean est lent à **comprendre**.
Le livre est difficile à **comprendre**.

Comme pour les verbes, c'est l'information lexicale sur les adjectifs *« lent »* et *« difficile »* qui détermine que le sujet du premier est le sujet de *« comprendre »*, alors que le sujet du second est l'objet direct de *« comprendre »*.

Les adjectifs à contrôle forment deux classes, selon que leur sujet est le sujet ou l'objet direct de l'infinitif contrôlé. Ces deux cas se traduisent chacun par une règle ; la règle `TOUGH-MOVEMENT` de la figure 1 correspond au cas où c'est l'objet de l'infinitif qui est contrôlé.

Il n'existe pas, pour les adjectifs, l'équivalent de Dicovalence pour les verbes. Nous avons donc relevé les adjectifs du corpus qui présentaient un argument infinitif et nous les avons classés manuellement pour ancrer les règles.

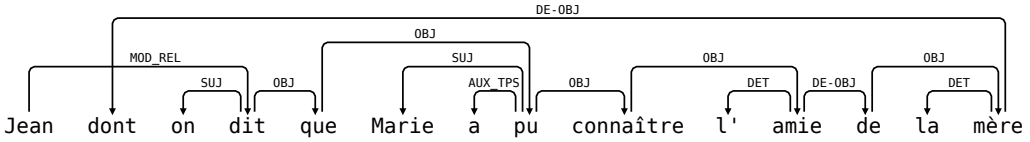
2.3 Les antécédents d'anaphores syntaxiques

Connaître les liens anaphoriques qui sont totalement déterminés par la syntaxe permet de calculer certaines co-références sémantiques. Dans ce travail, nous en distinguons trois types.

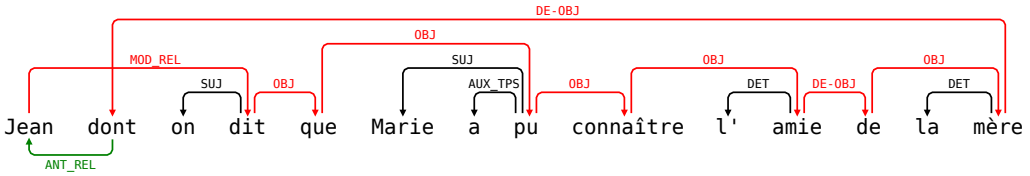
Le premier est le pronom réfléchi complément réel d'un verbe : son antécédent est le sujet de ce verbe. Par exemple, dans *« Jean ne veut pas* **se** *laver »*, l'objet direct de *« laver »* est *« se »*, qui co-réfère sémantiquement avec *« Jean »*. Une règle permet de détecter ce cas et une relation notée ANT_REFL entre le pronom et son antécédent est ajoutée.

Le second est le pronom personnel répétition d'un sujet : son antécédent est l'autre sujet du verbe. Dans *« Jean veut-***il** *manger ? »*, *« veut »* a deux sujets, dont *« il »* qui co-réfère sémantiquement avec *« Jean »*. Une relation ANT_REP est alors ajoutée.

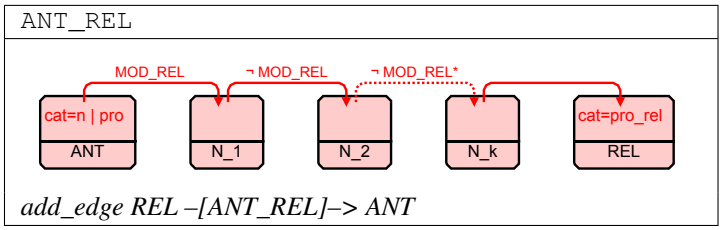
Le troisième type de lien anaphorique que nous considérons relie un pronom relatif à son antécédent. La création de ce lien, noté ANT_REL, est un problème plus complexe. Dans le FTB, il faut pour cela remonter les dépendances depuis le pronom relatif jusqu'à la tête de la relative, laquelle est reliée à l'antécédent par une dépendance MOD_REL. Considérons l'expression suivante annotée selon le guide du FTB :

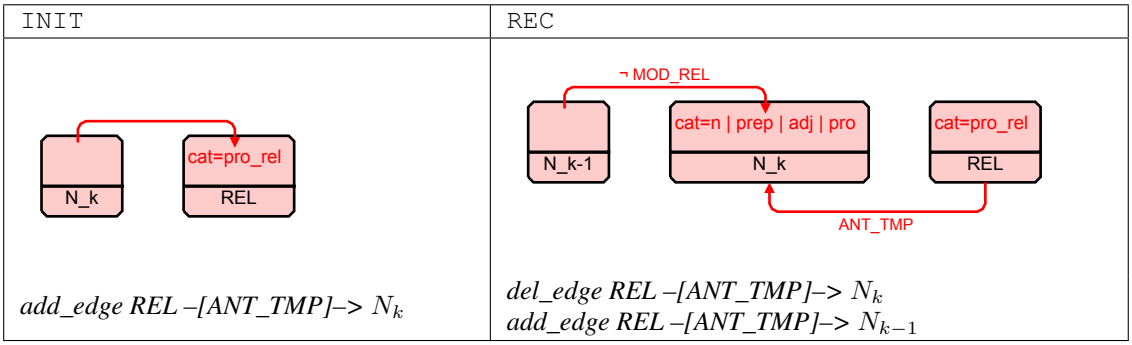


Dans cet exemple, on passe du pronom relatif à son antécédent en remontant successivement les dépendances DE-OBJ, OBJ, DE-OBJ, OBJ, OBJ, OBJ, OBJ, MOD_REL (en rouges dans la figure ci-dessous). Finalement, la nouvelle dépendance ANT_REL (en vert et en bas sur la figure), est produite.

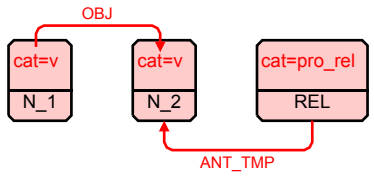
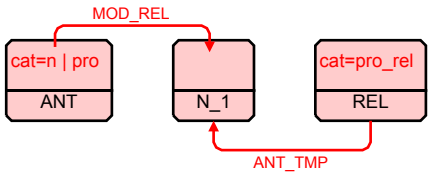


La transformation peut être décrite à l'aide du motif généralisé ci-contre. Toutefois, d'une part, ce type de motif ne permet pas d'exprimer les contraintes (comme les contraintes d'îlots) sur les structures de trait des noeuds. D'autre part, un tel motif n'est plus local, or cette propriété de localité du calcul permet une implantation efficace. La solution proposée ici emploie un lien temporaire ANT_TMP entre le pronom relatif et le point de la chaîne où on se situe dans sa remontée. Les quatre règles ci-dessous implémentent le motif généralisé en prenant en compte les contraintes d'îlots [6].





6. Des motifs négatifs (non représentés dans les figures) permettent de bloquer l'application d'une règle : par exemple, la règle INIT ne doit s'appliquer que sur un pronom relatif où il n'y a pas encore de lien ANT de créé.

RECV	STOP
	
del_edge REL –[ANT_TMP]–> N_2 *add_edge REL –[ANT_TMP]–> N_1*	*del_edge REL –[ANT_TMP]–> N_1* *add_edge REL –[ANT_REL]–> ANT*

Les relatives qui sont en apposition ou les constructions clivées peuvent être repérées par le même mécanisme de règle généralisée décomposée en règles atomiques (en changeant seulement la dernière règle). Les exemples suivants sont issus du corpus :

C'est une novice inspirée qui redressa [...]
Une mutation est en marche, qui ne s'arrêtera sans doute pas [...]

2.4 La coordination

La coordination est annotée dans le FTB de façon à respecter la contrainte d'arbre : un lien COORD va de la tête du premier conjoint jusqu'à la conjonction de coordination et un autre lien DEP_COORD va de la conjonction jusqu'à la tête du second conjoint. Si la tête du second conjoint est un verbe ou un adjectif, il est nécessaire de déterminer ses actants syntaxiques pour calculer sa représentation sémantique ; or les actants syntaxiques du second conjoint ne sont pas annotés dans le FTB quand ils sont partagés avec le premier conjoint. Considérons quelques exemples de coordination de syntagmes verbaux ou adjectivaux, où la tête du second conjoint est en gras et les actants partagés soulignés :

Jean déballe et **mange** son casse-croûte.
Abandonnée de son mari et **dégoûtée**, Marie ne sort plus.
Jean pense partir aujourd'hui et **pouvoir** rentrer dans un mois.

Dans le cadre de cet article, nous nous limitons à la recherche du sujet du second conjoint. Pour cela, on crée une dépendance SUJ entre le verbe tête du second conjoint et le sujet du premier conjoint, à condition que le second conjoint n'ait pas déjà un sujet. Deux règles différentes créent cette dépendance, selon que les syntagmes verbaux coordonnés sont introduits ou non par une préposition ou un complémenteur.

3 Mise en œuvre

3.1 Organisation en modules

Nous avons montré dans (Bonfante *et al.*, 2011) qu'une organisation modulaire permettait d'avoir un contrôle global sur le calcul et de simplifier le développement d'un système de règles. Les modules sont déterminés lin-

guistiquement : un module est un ensemble de règles contribuant à une transformation linguistique particulière qu'il est possible d'isoler relativement aux autres règles.

Cette organisation en modules permet également de gérer le comportement global du système de réécriture. Dans le cas présent, la terminaison est vérifiée pour chaque module, elle l'est donc globalement. Comme on l'a vu dans le cas des verbes à contrôle, notre enrichissement est nécessairement non-confluent mais cette non-confluence est très localisée (lexicalement) et donc contrôlable.

Par rapport à notre objectif général de construire une représentation sémantique à partir d'une représentation syntaxique en dépendances, les modules jouent un rôle essentiel. Par rapport à la tâche limitée qui est celle présentée dans cet article et qui consiste à enrichir les dépendances du FTB, ils ont une portée plus restreinte. Les interactions possibles entre les règles permettent difficilement de les isoler en modules.

Ainsi, les règles grammaticales d'actants et les règles lexicales d'actants peuvent se combiner de plusieurs façons. Prenons deux exemples. Dans la phrase *« Jean boit pour* **essayer** *d'***oublier** *»*, une règle grammaticale établit que *« Jean »* est sujet d'*« essayer »*, puis une règle lexicale ajoute que *« Jean »* est aussi sujet d'*« oublier »*. Dans la phrase *« Marie interdit à Jean de* **boire** *pour* **oublier** *»*, une règle lexicale établit *« Jean »* comme sujet de *« boire »*, puis une règle grammaticale établit que *« Jean »* est aussi sujet d'*« oublier »*.

Séparer les règles lexicales et grammaticales dans des modules différents ne permettrait pas de capturer toutes leurs interactions. C'est la raison pour laquelle nous répartissons les règles, au nombre de 47, présentées dans les sous-sections précédentes en trois modules ordonnés comme suit :
– le premier module ajoute les relations ANT des pronoms réfléchis, répétitions de sujets et relatifs vers leurs antécédents. Il contient 8 règles : 6 pour les relatives qui décomposent le motif généralisé, 1 pour les pronoms réfléchis et 1 pour les pronoms personnels répétitions de sujets ;
– le second module, SUJET, ajoute les relations SUJ des adjectifs, participes et infinitifs, que la règle soit grammaticale ou lexicale. Il contient 36 règles :11 règles grammaticales et 25 règles lexicales ;
– le troisième module, OBJET, ajoute les relations OBJ des infinitifs contrôlés dans des constructions de *tough movement*. Il contient 3 règles.

Le fait d'ajouter des sujets aux adjectifs permet de gérer le *tough movement* avec une seule règle. En revanche, il est alors nécessaire d'appliquer le module SUJET avant le module OBJET.

Enfin, la coordination intervient dans tous les modules et nécessite un traitement particulier. Reprenons un exemple déjà introduit pour la coordination :

Jean pense partir aujourd'hui et **pouvoir** rentrer dans un mois.

Dans cet exemple, il faut déterminer le sujet de *« partir »* avant d'en déduire que *« pouvoir »* partage ce sujet. Cela incite donc à appliquer les règles de coordination après le module SUJET. Cependant, pour déterminer le sujet de *« rentrer »*, il est nécessaire de connaître le sujet de *« pouvoir »*. De ce point de vue, il faudrait donc appliquer les règles relatives à la coordination avant le module SUJET.

Un moyen de résoudre cette contradiction est de particulariser chaque règle de coordination en autant de règles qu'il existe de modules concernés.

3.2 Expériences

Nous avons appliqué notre système aux 12 351 phrases annotées en dépendances que contient le FTB. Un extrait de 120 phrases est disponible en ligne [7].

7. http://wikilligramme.loria.fr/doku.php?id=taln2011

Les relations ANT, SUJ et OBJ que nous ajoutons sont très courantes dans le corpus : on en observe sur 87% des phrases. En outre, on ajoute, sur chaque phrase, en moyenne, trois nouvelles relations. On peut remarquer le faible nombre de relations OBJ ajoutées qui montrent que le *tough movement* est assez rare en corpus. Pour avoir une évaluation plus fine, notamment de la précision, il faudrait disposer d'une partie du corpus annotée manuellement qui n'existe malheureusement pas pour l'instant.

Au total, nous ajoutons :
– 3 691 relations ANT (3 152 pour les pronoms relatifs, 99 pour les pronoms réfléchis et 270 pour les pronoms personnels répétitions de sujets) ;
– 33 605 relations SUJ (23 940 pour les adjectifs et 9 665 pour les verbes) ;
– 19 relations OBJ.

Pour avoir une idée précise de la pertinence de ce système, il convient également d'essayer de détecter les cas problématiques, c'est-à-dire les cas où des relations sont susceptibles de manquer dans les structures produites par notre système de réécriture. Pour cela, nous avons observé les trois configurations suivantes :
– les verbes (sauf impératifs et auxiliaires) qui n'ont toujours pas de sujet (de surface ou profond) après réécriture, il y en a 3 548 ;
– les adjectifs sans sujet : 955
– les pronoms relatifs sans antécédent : 242

Pour chacune de ces configurations, nous avons étudié un échantillon de 100 phrases que nous avons classées manuellement. Le résultat de ce classement figure dans le tableau ci-dessous.

	Pas d'erreur	Annotation FTB	Problème de lexique	Problème de règles
verbes sans sujet	51	29	13	7
adjectifs sans sujet	32	45	1	22
pronoms relatifs sans antécédent	5	72	0	23

Les erreurs d'annotation du FTB que nous avons relevées sont systématiques et probablement liées à la conversion du FTB des constituants aux dépendances. L'annotation du FTB est donc la première source d'erreur dans notre processus. Les problèmes de lexique sont liés à des constructions nominales qui ne sont pas décrites dans Dicovalence. Nos règles ne couvrent pas certains phénomènes linguistiques : par exemple, un adjectif en emploi substantivé peut être antécédent d'une relative.

Discussion

(Gardent & Cerisara, 2010; Gardent, 2010) ont aussi utilisé la réécriture de graphes sur le FTB mais avec un objectif différent du nôtre, celui d'ajouter une annotation en rôles sémantiques pour les verbes. C'est pourquoi la plupart des règles qu'ils ont conçues vise à reformuler les constructions passives et causatives en constructions actives canoniques. Les quelques règles supplémentaires traitent de la coordination et des sujets des infinitifs. Les dernières, n'utilisant pas d'information lexicale, ont le défaut de systématiquement choisir pour sujet des infinitifs compléments celui du verbe dont ils dépendent. Enfin, l'étude n'aborde pas la question de la dépendance entre les règles et de leur ordre d'application.

Nous avons montré dans (Bonfante *et al.*, 2011) qu'il était possible de prendre compte les reformulations dans toute leur diversité (passif, moyen, causatif, impersonnel ...) à l'aide de la réécriture de graphes. Si nous avons écarté celles-ci du travail présenté dans cet article, c'est que nous nous étions fixé comme cadre pour ce travail d'enrichir les structures en dépendances du FTB sans modifier les relations existantes.

Intégrer les reformulations nécessiterait d'ajouter des règles correspondantes au système qui vient d'être présenté. Les interactions complexes entre ces règles nouvelles et les règles existantes font que cela ne peut pas se faire simplement par l'ajout d'un module avant ou après ceux qui ont été définis dans cet article. Par exemple, pour la phrase *« Jean est autorisé à* **partir** *»*, la reformulation du passif doit être effectuée avant la détermination du sujet profond de *« partir »*. C'est le contraire pour la phrase *« Jean demande à Marie d'être* **aidée** *de Pierre. »* En conséquence, l'intégration des reformulations devrait nous amener à restructurer le système de modules.

Nos expérimentations ont mis en évidence certaines limites du format d'annotation du FTB, qui sont dues essentiellement au choix des annotateurs de n'avoir que des arbres comme structures de dépendance. Premièrement, l'annotation des constructions causatives ne distingue pas le sujet et l'objet du verbe causé. Dans *« Jean fait* **manger** *un lapin »*, *« lapin »* est annoté OBJ de *« manger »*, que le lapin mange ou soit mangé. Deuxièmement, lorsque deux verbes sont coordonnés, l'annotation ne permet pas de retrouver les compléments partagés du second conjoint. Ainsi *« Jean* **déballe** *et* **mange** *son sandwich »* et *« Jean* **déballe** *son sandwich et* **mange** *»* ont la même annotation. Rien ne permet de distinguer que dans la première phrase, *« sandwich »* est également objet de *« déballe »* alors que dans la deuxième, *« sandwich »* n'est pas objet de *« mange »*.

Remerciements Nous remercions Sylvain Kahane pour une discussion qui a déclenché ce travail. Nous tenons également à remercier Anne Abeillé pour nous avoir autorisé à publier un extrait du FTB.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In *Treebanks. Building and Using Parsed Corpora*, chapter 10. Kluwer Academic Publishers.

BONFANTE G., GUILLAUME B., MOREY M. & PERRIER G. (2010). Réécriture de graphes de dépendances pour l'interface syntaxe-sémantique. In *TALN 2010*, Montréal, Canada.

BONFANTE G., GUILLAUME B., MOREY M. & PERRIER G. (2011). Modular graph rewriting to compute semantics. In *IWCS 2011*, p. 65–74, Oxford, UK.

CANDITO M.-H., CRABBÉ B., DENIS P. & GUÉRIN F. (2009). Analyse syntaxique statistique du français : des constituants aux dépendances. In *TALN 2009*, Senlis, France.

COPESTAKE A. (2009). *Invited Talk :* Slacker semantics : Why superficiality, dependency and avoidance of commitment can be the right way to go. In *EACL 2009*, Athens, Greece.

GARDENT C. (2010). Extraction des cadres syntaxiques à partir de P7dep. Notes transmises par l'auteur.

GARDENT C. & CERISARA C. (2010). Semi-Automatic Probanking for French. In *TLT9 – the ninth international workshop on Treebanks and Linguistic Theories, Tartu, Estonia*.

HAJIČ J., BÖHMOVÁ A., HAJIČOVÁ E. & HLADKÁ B. (2000). *The Prague Dependency Treebank : A Three-Level Annotation Scenario*, In *Treebanks : Building and Using Parsed Corpora*, p. 103–127. Amsterdam :Kluwer.

KAHANE S. (2001). Grammaires de dépendance formelles et théorie Sens-Texte. In *Tutoriel, TALN 2001*, volume 2, Tours.

REZAC M. (2006). *On tough-movement*, In *Minimalist Essays*, p. 288–325. Linguistik Aktuell/Linguistics Today 91. John Benjamins.

VAN DEN EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *French Language Studies*, **13**, 63–104.

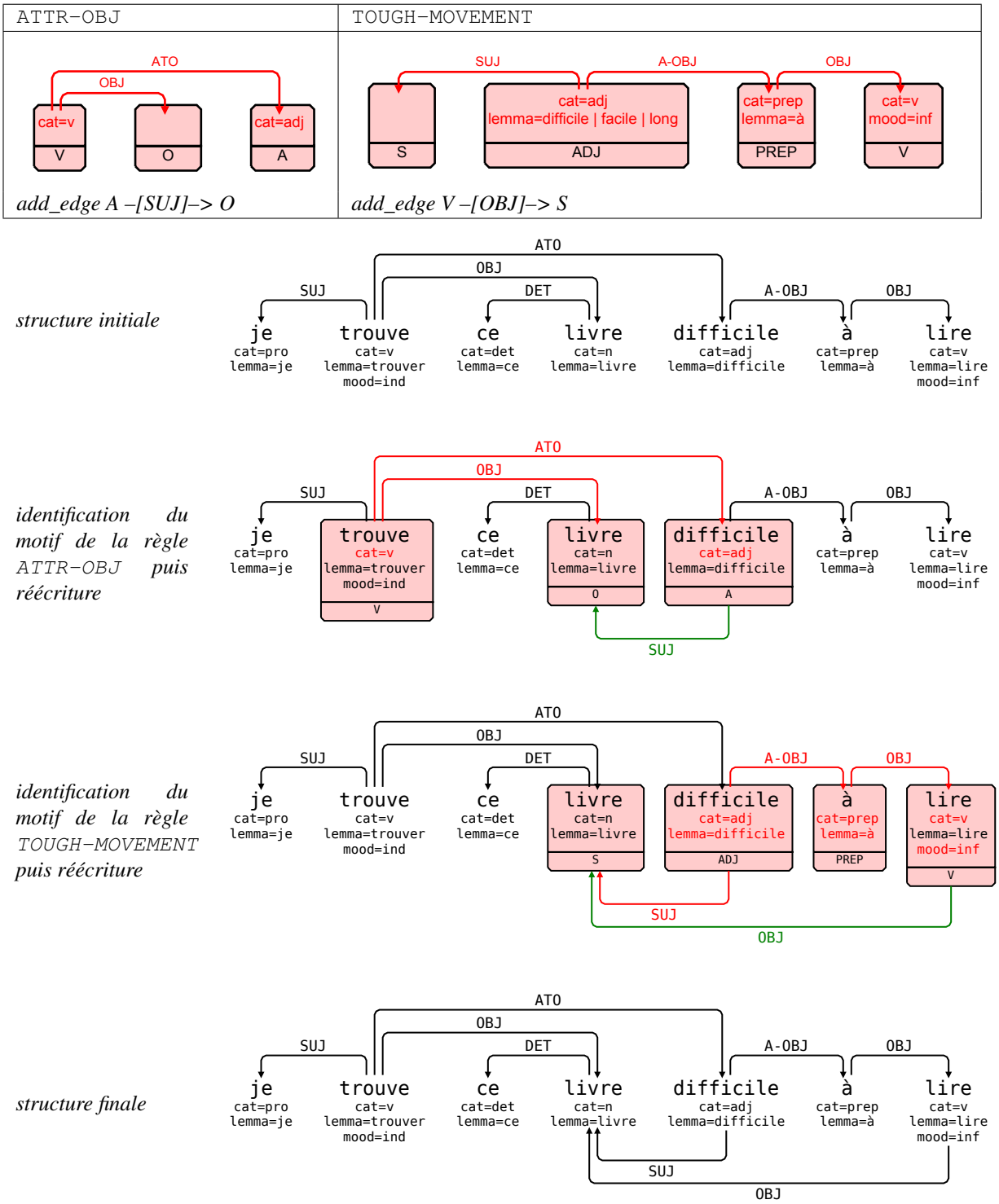


FIGURE 1 – *« Je trouve ce livre difficile à lire. »*