Enrichissement du FTB : un treebank hybride constituants/propriétés

Philippe Blache & Stéphane Rauzy
LPL, 5 Avenue Pasteur, 13100 Aix-en-Provence
{blache;rauzy}@lpl-aix.fr

RÉSUMÉ

Cet article présente les mécanismes de création d'un treebank hybride enrichissant le *FTB* à l'aide d'annotations dans le formalisme des *Grammaires de Propriétés*. Ce processus consiste à acquérir une grammaire *GP* à partir du treebank source et générer automatiquement les structures syntaxiques dans le formalisme cible en s'appuyant sur la spécification d'un schéma d'encodage adapté. Le résultat produit, en partant d'une version du *FTB* corrigée et modifiée en fonction de nos besoins, constitue une ressource ouvrant de nouvelles perspectives pour le traitement et la description du français.

ABSTRACT

Enriching the French Treebank with Properties

We present in this paper the hybridation of the French Treebank with Property Grammars annotations. This process consists in acquiring a PG grammar from the source treebank and generating the new syntactic encoding on top of the original one. The result is a new resource for French, opening the way to new tools and descriptions.

MOTS-CLÉS : Treebank hybride, French Treebank, Grammaires de Propriétés.

KEYWORDS: Hybrid treebank, French Treebank, Property Grammars.

1 Introduction

La constitution d'un treebank pour le français reste une priorité pour la description de notre langue. Il n'existe à ce jour quasiment qu'une seule ressource de ce type : le *French Treebank* (Abeillé *et al.*, 2003) à partir duquel quelques travaux ont été proposés (voir par exemple (Candito *et al.*, 2010), (Pynte *et al.*, 2001)). D'autres projets sont actuellement en cours dans différents laboratoires (voir (Cerisara *et al.*, 2010)), mais aucun ne propose à ce jour de véritable distribution. L'enrichissement du *FTB* est donc nécessaire pour disposer d'une ressource aussi complète que possible à partir de laquelle différents outils peuvent être développés. Parmi ces enrichissements, la constitution d'un treebank hybride comportant des analyses dans différents formalismes (voir par exemple (Candito *et al.*, 2010)) s'avère extrêmement utile non seulement d'un point de vue théorique (comparaison entre les différentes analyses), mais également pratique (notamment en termes d'apprentissage et d'évaluation). Par ailleurs, l'hybridation d'un treebank constitue un procédé économique pour la constitution d'une nouvelle ressource dans le formalisme cible : le fait de disposer d'une structure syntaxique (élaborée et vérifiée

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 307–320,
Grenoble, 4 au 8 juin 2012.

manuellement dans le cas du *FTB*) permet de générer automatiquement et de façon très contrôlée les structures dans le formalisme cible, héritant au passage des qualités du treebank d'origine.

Nous décrivons dans cet article l'enrichissement du *FTB* permettant la construction d'un treebank hybride *Grammaire Syntagmatique / Grammaire de Propriétés* . Dans une première partie, sans revenir sur une présentation du *FTB* disponible ailleurs, nous décrirons les corrections que nous avons effectuées (essentiellement des erreurs d'étiquetage) ainsi que les modifications apportées au format d'origine de façon à faciliter l'hybridation. Dans une seconde partie, nous proposons un schéma abstrait d'annotation pour les *Grammaires de Propriétés* (Blache, 2001) et son encodage en XML. La troisième partie sera consacrée à la présentation du procédé d'acquisition à partir du *FTB* de la grammaire dans le formalisme des Grammaires de Propriétés avant de décrire sa mise en oeuvre pour l'enrichissement du treebank. La dernière partie sera consacrée à une présentation et une discussion des résultats.

2 Le treebank FTB_{LPL}

Nous avons pour cette étude travaillé sur un sous-ensemble du *FTB*, que nous noterons désormais FTB_{LPL}(pour *LPL French Treebank*). Il a été constitué à partir du corpus *MFT* (*Modified French Treebank*, voir (Schluter et van Genabith, 2007)), lui-même sous-ensemble du corpus *FTB*.

Nous avons choisi d'apporter quelques modifications au format d'origine du *FTB* de façon à assurer une meilleure homogénéité avec les ressources existantes dans d'autres langues ou pour d'autres domaines (par exemple description de la multimodalité, études psycholinguistique, etc.). Ces modifications portent d'une part sur le niveau morphosyntaxique, pour lequel nous avons adopté le jeu de traits Multext (Ide et Véronis, 1994).

Dans un premier temps, l'étiqueteur du LPL (un étiqueteur stochastique basé sur le modèle des patrons (Blache et Rauzy, 2008; Rauzy et Blache, 2009) qui dans sa version actuelle atteint un score de F-Mesure de 0.975) a été appliqué à l'ensemble du corpus. Une fouille d'erreurs a ensuite été effectuée par correction ou validation manuelles des passages pour lesquels notre étiquetage présentait une différence avec celui proposé dans le *MFT*. D'autre part nous avons opéré une régularisation des positions des marqueurs de ponctuation, en déplaçant ces marqueurs à l'extérieur des syntagmes composant l'arbre.

Sur le plan syntaxique, nous avons systématisé certaines représentations (par exemple la projection des têtes) ou encore introduit de nouvelles catégories (plus conformes aux ressources existantes). Ces choix ne remettent pas en question la structure syntaxique d'origine, ils concernent essentiellement la forme.

Ces modifications ont été appliquées sur les 4.741 phrases du MFT (soit 134.445 mots). La moitié d'entre elles ont ensuite été validées manuellement. Environ 30% des phrases de ce sous-échantillon ont été temporairement rejetées faute d'un consensus clair dans la description de leur structure syntaxique ou si l'arbre proposé présentait des erreurs de catégorisation ou de rattachement. Le sous-ensemble du FTB_{LPL}utilisé dans cette étude compte finalement 1.471 phrases validées manuellement (soit environ 26.000 mots). La validation de la totalité du *MFT* est en cours.

A titre d'information, le schéma de la figure 1 récapitule le nombre d'occurrences des catégories

Catégories	*Modifications dans le FTB_{LPL}*
AP, PP, AdP	- Projection des têtes unaires
NP	- Projection des clitiques
	- Le N épithète fait partie du NP ("une tarte *maison*")
	- Plusieurs AP possibles dans le NP ("un *très bon premier* ministre")
Srel	- ProRel est directement rattaché à Srel (pas au NP ou PP de la relative)
VN	- Tous les participes se projettent en VN (sauf les participes passés des temps composés)
Coord :	- Projection d'un noeud COORD indiquant le type et la fonction des conjoints
VP	- Projection systématique, avec pour tête VN, incluant compléments et adjoints

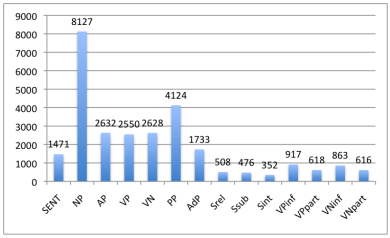

FIGURE 1 – Répartition des catégories dans le corpus

réalisées dans le FTB_{LPL}, indiquant une fréquence nettement supérieure du NP et dans une moindre mesure du PP par rapport aux autres catégories.

3 Un schéma d'annotation pour les *Grammaires de Propriétés*

L'hybridation du *FTB* consiste à ajouter aux données d'origine les annotations syntaxiques issues de l'analyse en *Grammaires de Propriétés* (voir (Blache, 2001)). Il s'agit d'une approche reposant sur la notion de construction (à la manière des *Grammaires de Construction* (voir (Kay et Fillmore, 1999))), dans laquelle les unités de base sont des objets (également appelés *signes*) comportant certaines caractéristiques décrites par des traits. En *Grammaires de Propriétés* (notées dorénavant *GP*), les signes ne sont porteurs que d'informations statiques (ou traits endogènes). Pour les signes lexicaux, il s'agit par exemple des traits morpho-syntaxiques, de la forme, sa position dans la phrase, etc. Ces informations sont récapitulées dans la structure de traits de la figure 3.

Les signes entretiennent entre eux des relations appelées *propriétés*. Les *GP* reposant sur une représentation explicite de toutes les informations syntaxiques, chaque type d'information correspond ainsi à un type de propriété qui sont, dans la grammaire élaborée pour ce treebank, au nombre de six :

– *Linéarité* : relations de précédence linéaire
– *Obligation* : identification de la tête du syntagme
– *Dépendance* : relation syntactico-sémantique entre les catégories
– *Unicité* : catégories qui dans un syntagme ne peuvent être répétées

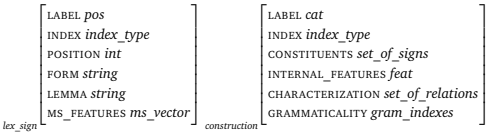

FIGURE 2 – Structures des signes lexicaux et des constructions

– *Exigence* : cooccurrence obligatoire de catégories
– *Exclusion* : impossibilité de cooccurrence de catégories

A ces types de propriétés s'ajoute la spécification de l'ensemble des constituants pouvant intervenir dans la réalisation d'un syntagme. Notons que l'ensemble de propriétés décrit ici n'est pas limitatif, d'autres types d'informations peuvent également être ajoutés, comme l'adjacence ou la représentation explicite de l'accord. Au total, un syntagme sera décrit par un couple *<constituants, propriétés>* comportant d'une part la liste de ses constituants possibles et d'autre part l'ensemble des propriétés qui forment des relations entre les constituants (ou dans le cas des propriétés unaires comme l'obligation ou l'unicité, entre le constituant et la racine).

Dans le vocabulaire des *GP*, les relations unissant les constituants d'un syntagme forment l'ensemble des propriétés évaluées, appelé *caractérisation*. Un ensemble de signes reliés par des propriétés correspond dans notre approche à la description d'une *construction*. D'une façon générale, en *GP*, toutes les unités de niveau non lexical, et notamment les unités syntaxiques, sont considérées comme des constructions. De plus, de la même façon que pour les signes lexicaux, une construction peut posséder des informations intrinsèques, décrites de façon statique par des traits spécifiques (par exemple, des informations morpho-syntaxiques ou grammaticales). Une construction correspond donc à la seconde structure de la figure 3.

Notons qu'un des traits spécifiques de la construction concerne la grammaticalité. Il s'agit d'un ensemble d'indices permettant de décrire le niveau de grammaticalité (Blache *et al.*, 2006). Ces indices s'appuient notamment sur la mesure de la densité des relations (nombre de propriétés satisfaites), leur importance (poids des propriétés) etc. à partir desquels un indice global de grammaticalité peut être proposé. L'indice de grammaticalité constitue une information intéressante, y compris dans le cas de constructions "bien formées" comme dans le cas du FTB. Cet indice permet en effet de mesurer la densité d'information syntaxique construite et constitue un élément d'évaluation de la complexité de la structure, utile notamment dans la perspective d'expériences en psycholinguistique.

La description d'une construction forme un *graphe de description* dans lequel les signes et les propriétés sont respectivement représentés par des nœuds et des arcs. Dans cette représentation, toute construction correspond à un graphe dont la racine correspond à l'identification de cette construction : une construction peut à son tour être utilisée comme constituant d'une autre construction. Les signes, correspondant aux nœuds du graphe, sont donc aussi bien des objets lexicaux que des constructions.

L'exemple de la figure 3 propose le graphe de description pour une phrase extraite du FTB_{LPL}. les nœuds du graphe représentent des signes (pour des raisons de simplicité, seuls quelques traits sont représentés), tandis que les arcs portent les différentes propriétés : précédence, dépendance,

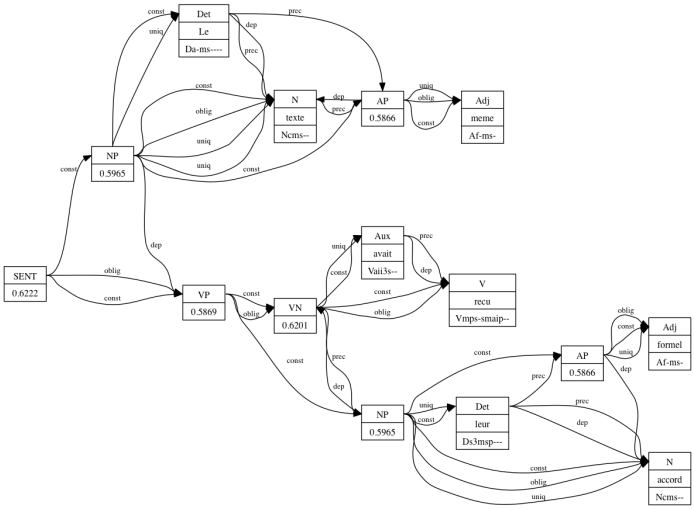

FIGURE 3 – Graphe de description de la phrase "*Le texte même avait reçu leur accord formel*"

obligation, constituance, unicité. Notons que les relations d'exclusion (qui jouent surtout un rôle en cas de violation de contrainte) ne sont pas indiquées là encore pour des raisons de lisibilité.

En *GP*, toutes les relations sont indiquées au même niveau. En isolant les sous-graphes de certaines de ces relations, il est possible d'extraire du graphe un arbre syntagmatique (sous-graphe formé des arcs étiquetés *const*) ou de dépendance (arcs étiquetés *dep*).

Dans le cadre du FTB_{LPL}, s'agissant de données écrites, les constructions concernent seulement le domaine syntaxique : chaque syntagme correspond à une construction définie par l'ensemble des nœuds participants (les constituants) et l'ensemble des relations qui les relient.

4 Schéma d'encodage XML

Le schéma abstrait défini dans la section précédente permet de définir un schéma d'encodage XML. Les signes sont tout d'abord caractérisés par leur type (signe lexical ou construction) qui induira des structures différentes pour les structures de traits qui les décrivent. Tous les types ont les attributs suivants :

< `sign` >

type	type du signe (lexical ou construction)
label	identification de la catégorie
index	index du noeud permettant sa référence
features	vecteur de traits morpho-syntaxiques, traits syntaxiques

Les signes lexicaux sont décrits par les attributs suivants :

Type *lex*

position	identification de la position dans la chaîne
form	forme du mot
lemma	lemme

De leur côté, les signes de type *construction* dans l'encodage *GP* sont décrits par un ensemble d'attributs portant les informations intrinsèques ainsi que l'indice du syntagme correspondant dans l'encodage *FTB* :

Type *const*

label	identification de la catégorie
index	index du nœud permettant sa référence
ftb_index	index du nœud correspondant dans l'encodage *FTB*

Par ailleurs, un signe de type *construction* contient un certain nombre d'éléments décrivant l'ensemble de constituants, leurs propriétés et l'évaluation de leur grammaticalité :

Type *const* : :=

< `constituents` >	ensemble des signes participant à la construction
< `characterization` >	ensemble des relations (propriétés) entre les signes
< `gram_indexes` >	ensemble des indices d'évaluation de la grammaticalité

La caractérisation d'une construction est formée quant à elle par l'ensemble des propriétés. Chaque propriété est décrite par un ensemble d'attributs décrivant le type de la relation, les nœuds sources et cibles ainsi que la satisfaction (ou violation) de la propriété :

< `characterization` >

type	type de la propriété (linéarité, dépendance, etc.)
source	index du nœud source de la relation (nœud racine en cas de relation unaire)
target	index du nœud cible
sat	satisfaction de la propriété (valeur '*plus*' ou '*moins*')

L'exemple suivant illustre l'encodage du dernier syntagme nominal de la phrase "*Le texte même avait reçu leur accord formel*" dont le graphe de description est donné plus haut. La première partie de l'exemple fournit l'encodage des annotations sous format FTB modifié comme indiqué en première section. Il s'agit d'une représentation arborescente classique, dans laquelle toutes les catégories, lexicales et syntagmatiques, sont représentées par des éléments de type `sign`.

Encodage FTB

```
<sign type="const" label="NP" features="NP_OBJ" index="f-10">
     <sign type="lex" label="Det" features="Ds3msp--" index="f-11" form="leur" lemma="leur"/>
     <sign type="lex" label="N" features="Ncms-" index="f-12" form="accord" lemma="accord"/>
     <sign type="const" label="AP" features="AP" index="f-13">
          <sign type="const" label="Adj" features="Af-ms-" index="f-14" form="formel" lemma="formel"/>
     </sign>
</sign>
```

L'encodage en *GP* suit le schéma abstrait décrit plus haut. Un graphe de description est un ensemble de constructions, chacune formant un graphe composé d'un ensemble de nœuds (les constituants) et d'un ensemble de relations indiquées dans la caractérisation et constituant le cœur de la description syntaxique. Dans cet exemple, pour des raisons de lisibilité, elles sont représentées sous forme compacte pour la représentation de la caractérisation du *NP* (les propriétés sont des attributs dont les valeurs sont des couples d'index). Par ailleurs, le treebank étant hybride, les deux encodages *GP* et *FTB* sont représentés simultanément. Deux options sont

possibles pour l'encodage XML : la première consistant à représenter dans un fichier unique (donc à l'intérieur d'un arbre XML) les informations d'origine et les propriétés. La seconde option, plus fidèle à l'approche des GP repose sur une représentation parallèle : l'arbre XML correspondant au FTB d'un côté et les informations GP de l'autre. L'identification des signes lexicaux dans la partie d'encodage *GP* se fait donc par référence à l'index des éléments correspondants décrits dans l'encodage *FTB*.

Encodage GP

```
<sign type="const" label="AP" index="g-15" ftb_index="f-13">
     <constituents>
          <constituent sign_index="f-14"/>
     </constituents>
     <characterization>
          <property type="oblig" source="f-13" target="f-14" sat="p"/>
          <property type="unic" source="f-13" target="f-14" sat="p"/>
     </characterization>
     <gram_indexes gram="0.585" sat="1.0" complete="0.117" quality="1.0" precision="0.584"/>
</sign>
<sign type="const" label="NP" index="g-16" ftb_index="f-10">
     <constituents>
          <constituent sign_index="f-11"/>
          <constituent sign_index="f-12"/>
          <constituent sign_index="g-15"/>
     </constituents>
     <characterization prec_p="11:12;11:13" oblig_p="12" dep_p="11:12;13:12"
          exig_p="12:11;13:12" unic_p="11;12"/>
     <gram_indexes gram="0.595" sat="1.0" complete="0.145" quality="1.0" precision="0.597"/>
</sign>
```

Il est important de préciser que dans l'encodage *GP*, les constructions sont toutes représentées au même niveau : les relations entre les différents signes (les arcs du graphe) sont données par les propriétés et ne suivent pas nécessairement une hiérarchie stricte. En d'autres termes, les informations linguistiques sont représentées sous forme de graphe, et non d'arbre. Ce type de représentation est indispensable dans la perspective d'intégration d'informations issues de domaines autres que la syntaxe (discours, pragmatique ou encore, pour les corpus oraux, prosodie, phonétique, etc.). Son intérêt réside en particulier dans sa capacité à décrire des phénomènes discontinus, fréquents à l'oral, mais également à l'écrit.

Concrètement, l'encodage proposé suit ainsi les recommandations du format *GrAF* (voir (Ide et Suderman, 2007), (ISO24612, 2008)) : les différents éléments du graphe sont représentés sous la forme d'éléments non hiérarchisés, leur structuration étant fournie par les relations spécifiées entre les index. Cependant, à la différence de *GrAF*, nous avons choisi de représenter les informations statiques sous la forme d'attributs (lorsqu'il s'agit de vecteurs de traits) ou d'éléments (pour les structures de traits) à l'intérieur des nœuds. Dans notre représentation, les éléments `<sign>` correspondent aux nœuds. Ils peuvent faire référence au signe correspondant dans l'encodage *FTB* (attribut *ftb_index*). Ils comportent par ailleurs la spécification de l'ensemble des nœuds qui vont former le graphe (éléments `<constituents>`). La liste des relations (éléments `<property>`) est indiquée dans l'élément `<characterization>`. Elle est complétée par un élément portant les indices de grammaticalité.

Cette représentation permet d'identifier les constructions comme des unités à part entière et qui peuvent à leur tour devenir des nœuds du graphe de description. Ce dernier correspond donc à un *hypergraphe* dans lequel chaque nœud correspond soit à un élément atomique, soit à un

graphe. Dans notre exemple, la construction *AP* (index *g-15*) est ainsi décrite par un graphe qui est utilisé comme nœud dans la construction *NP* (index *g-16*).

5 Acquisition d'une GP à partir du FTB_{LPL}

La construction de la grammaire *GP* acquise à partir du FTB_{LPL}a été faite manuellement. Un outil de navigation et d'édition du FTB_{LPL}a pour cela été développé. Sa première fonctionnalité permet de lister pour chaque syntagme toutes ses réalisations possibles. Il s'agit en d'autres termes d'identifier, dans le cadre d'une grammaire syntagmatique, toutes les parties droites de règles. Le résultat se présente sous la forme d'un tableau HTML (voir l'exemple de la figure 4) permettant d'éditer pour chaque syntagme ses réalisations dans le corpus. La colonne de droite est constituée de liens vers les positions correspondantes dans le FTB_{LPL}, permettant ainsi de visualiser les exemples comme représenté dans la figure 5.

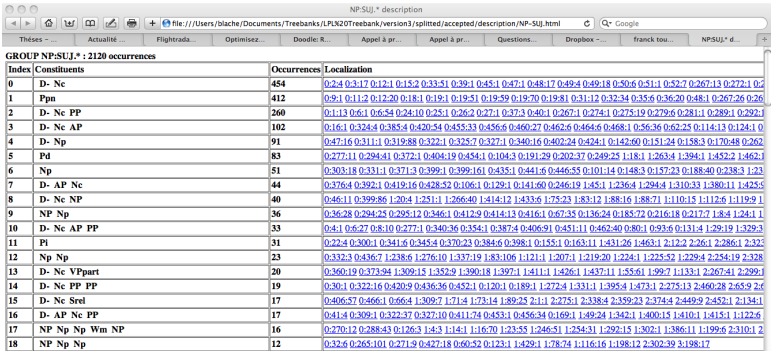

FIGURE 4 – Edition des constructions du NP sujet

Nous avons répertorié, grâce à l'éditeur ci-dessus, l'ensemble des constructions possibles de toutes les unités syntaxiques. Cette base d'information fournit directement la liste des constituants et leurs propriétés. Celles-ci sont construites de la façon suivante :

- *Linéarité* : pour chaque constituant, toutes les catégories pouvant la précéder, mais pas la suivre
- *Obligation* : liste des catégories, mutuellement exclusives, apparaissant obligatoirement dans une construction du syntagme. Il s'agit des têtes, il peut y en avoir plusieurs pour un même syntagme (mais jamais réalisées simultanément)
- *Unicité* : liste des catégories n'apparaissant qu'une fois dans chacune des constructions
- *Exigence* : pour chaque constituant, liste des catégories du syntagme apparaissant toujours avec lui
- *Exclusion* : pour chaque constituant, liste des catégories du syntagme n'apparaissant jamais avec lui

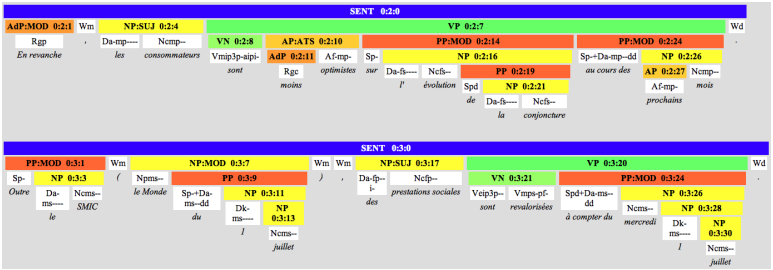

FIGURE 5 – Visualisation d'arbres

Remarquons que ces propriétés peuvent être acquises automatiquement à partir de la liste des constructions possibles du syntagme sur la base de laquelle les informations sont extraites. A ces propriétés s'ajoute la relation de *dépendance*. Celle-ci correspond, dans le cas de cette grammaire, aux relations de complémentation, adjonction et spécification. Il est toujours possible d'enrichir ou de modifier cet ensemble de relations. Le tableau suivant récapitule pour chaque catégorie le nombre des propriétés acquise à partir du *FTB* :

	const	lin	dep	unic	oblig	exig	excl
SENT	8	5	3	3	1	0	2
NP	12	36	18	10	4	6	44
AP	5	7	4	3	1	0	2
PP	7	6	5	3	1	0	0
AdP	10	18	5	4	1	0	1
VP	10	24	8	3	1	0	0
VPinf	9	13	0	7	1	0	0
VPpart	8	7	0	7	1	0	0
VN	6	11	5	2	1	0	0
VNinf	7	6	5	4	3	4	0
VNpart	7	7	6	5	2	4	0
Srel	7	7	0	5	1	0	3
Ssub	10	14	1	1	1	0	3
Sint	8	4	0	5	1	0	6

Ces propriétés sont inégalement réparties, le *NP* étant la catégorie la plus riche. On constate par ailleurs que le nombre de propriétés et leur répartition entre les types ne sont pas totalement dépendants du nombre de catégories, mais plutôt de la variété des constructions possibles. La figure 6 propose l'exemple de l'ensemble des propriétés du *NP* observées dans le FTB_{LPL}.

6 Enrichissement automatique du FTB_{LPL}

L'enrichissement du FTB_{LPL}par une description en *GP* se fait automatiquement. Nous utilisons pour cela un ensemble de solveurs de contraintes appliqué à l'entrée sous format XML du FTB_{LPL}. Le processus d'évaluation des contraintes permettant de construire l'analyse en *GP* est donc très fortement contraint : il prend en effet en entrée un syntagme déjà identifié, ainsi que la liste de ses constituants. Ceci permet donc de sélectionner dans la grammaire le sous-ensemble

const	Det, N, N_p, Pro, Clit, AdP, AP, NP, VPpart, VPinf, Srel, PP, Ssub
lin	Det $\prec$ {N, N_p, AdP, AP, VPpart, VPinf, Ssub, Srel, PP, NP} N $\prec$ {AdP, VPpart, VPinf, Ssub, Srel, PP, NP} N_p $\prec$ {AP, VPpart, Srel, PP} AP $\prec$ {VPpart, VPinf, Ssub, Srel, NP} AdP $\prec$ {Ssub, Srel, PP} NP $\prec$ {PP} Pro $\prec$ {Srel, PP} PP $\prec$ {VPpart, VPinf, Srel}
dep	{Det, AP, AdP, NP, VPpart, VPinf, PP, Srel, Ssub} $\rightsquigarrow$ N {AP, NP, VPpart, PP, Srel} $\rightsquigarrow$ N_p {PP, Srel} $\rightsquigarrow$ Pro
unic	unic = {Det, N, Pro, Clit, AdP, NP, VPpart, VPinf, Srel, Ssub}
oblig	oblig = {N, N_p, Pro, Clit}
exig	{Det, PP, AP, VPpart, VPinf, Ssub} $\Rightarrow$ N
excl	Pro $\otimes$ {Det, N, N_p, AP, AdP, NP, VPpart, VPinf, Ssub} AdP $\otimes$ {VPpart, VPinf} Ssub $\otimes$ {AP, NP, VPpart, VPinf, Srel, PP} NP $\otimes$ {Pro, VPpart, VPinf, Srel, Ssub} VPpart $\otimes$ {VPinf, Srel, Ssub} VPinf $\otimes$ {AP, Srel, Ssub} Srel $\otimes$ {Ssub} N_p $\otimes$ {N, Pro, AdP, VPinf, Ssub} Clit $\otimes$ {Det, N, Pro, AP, AdP, NP, VPpart, VPinf, Ssub}

FIGURE 6 – Propriétés du NP

de contraintes décrivant le syntagme en question là où un processus normal d'analyse en *GP* consisterait à parcourir l'ensemble des contraintes de la grammaire. Au total, le principe de construction de la description en *GP* consiste donc à parcourir le treebank initial et calculer la caractérisation de chaque syntagme du corpus. Celle-ci, comme indiqué plus haut, est formée d'une part par la liste des constituants (qui formera la liste de nœuds du graphe de description) et la liste des relations qui les relient. Cette dernière est obtenue en appliquant les solveurs de contraintes. La description de leur implantation peut être décrite de façon simplifiée par une présentation ensembliste.

On note : $|E|$ la cardinalité de l'ensemble E ; $\mathscr{C}$ la suite ordonnée des constituants du syntagme analysé, $\mathscr{C}_{i..j}$ le sous-ensemble de $\mathscr{C}$ entre les positions i et j ; c_i un constituant de $\mathscr{C}$ à la position i ; n le nombre de constituants de $\mathscr{C}$.

Par ailleurs, les propriétés dans la grammaire sont représentées soit par des ensembles (obligation, unicité) soit par des relations binaires. Dans le premier cas, on note $\mathscr{P}$ les catégories spécifiées dans l'ensemble et p_i une propriété de $\mathscr{P}$. Dans le second cas, on note *left* la catégorie de la partie gauche de la relation et *right* celle de la partie droite.

Obligation	$\lvert\mathscr{C} \cap \mathscr{P}\rvert > 0$
Linéarité	$left \in \mathscr{C}_{1..i} \Rightarrow right \in \mathscr{C}_{i+1..n}$
Unicité	$\forall c_i \in p_i, \lvert\{c_i\} \cap \mathscr{C}\rvert = 1$
Exigence	$left \in \mathscr{C} \Rightarrow right \in \mathscr{C}$
Exclusion	$left \in \mathscr{C} \Rightarrow right \notin \mathscr{C}$

Les descriptions opérationnelles ci-dessus sont décrites pour une propriété identifiée. La construction de la caractérisation totale consiste à appliquer cette évaluation à l'ensemble des propriétés décrivant le syntagme. Concrètement, nous avons développé un système couplant à un analyseur syntaxique probabiliste (entraîné sur le *FTB*) l'évaluateur *GP* développé comme indiqué

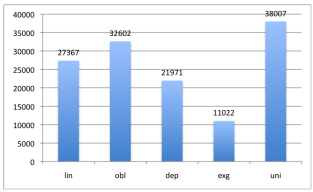

FIGURE 7 – Répartition des propriétés dans le corpus

ci-dessus. Nous pouvons donc désormais générer automatiquement un treebank hybride constituants/propriétés au format FTB_{LPL}.

7 Analyse de la répartition des propriétés

La génération des analyses en propriétés du *FTB* permet de préciser la distribution des propriétés de façon à en mesurer l'impact. L'ensemble des caractérisations permet en effet d'identifier les propriétés effectivement utilisées pour la description du corpus (on parle alors de propriétés *pertinentes*). L'étude de la distribution de ces propriétés fournit des indications précises au niveau général sur le rôle des différents types de propriétés dans la grammaire, mais également sur l'impact spécifique d'une propriété particulière dans la description d'un syntagme. Il devient donc possible d'envisager l'identification sur la base de corpus d'une répartition entre propriétés fortes et faibles, comme cela est proposé dans les approches en psycholinguistique [1].

La figure 7 indique la répartition du nombre total des propriétés pertinentes pour la description des unités du FTB_{LPL}. Nous avons isolé des indicateurs la propriété d'exclusion, celle-ci étant en effet inégalement répartie entre les syntagmes (beaucoup n'en contiennent pas). De plus, ce type de propriété consiste à vérifier l'absence de certaines catégories (restriction de cooccurrence), elle est donc presque toujours pertinente (ou évaluée), ce qui n'est pas le cas des autres propriétés. Enfin, cette propriété est directement dépendante du nombre de catégories pouvant être réalisées comme constituant d'un syntagme. D'une façon générale, les résultats montrent une fréquence importante pour les contraintes d'*unicité*, d'*obligation* et de *linéarité*. Les propriétés de *dépendance* et d'*exigence* sont quant à elles moins fréquentes dans les caractérisations. Cette information souligne l'importance de trois types d'informations non présentes explicitement dans une représentation syntagmatique classique : la présence de la tête, l'impossibilité de répétition d'un élément et l'ordre linéaire. En étudiant plus précisément la répartition de ces types de propriétés dans les syntagmes (voir figure 8), on constate que les catégories *SENT, NP* et dans une moindre mesure *PP* mobilisent l'essentiel des propriétés évaluées. Ce phénomène révèle en fait un certain degré de figement des constructions : les syntagmes ayant une grande variabilité notamment en termes de nombre de constituants et de variété des constructions ont recours à un plus grand nombre de propriétés pour leur description que les autres. Cette tendance se confirme en étudiant la répartition des propriétés évaluées pour les syntagmes montrant des description

1. Une telle répartition en deux types est plus opérationnelle pour le type d'analyse que nous proposons qu'une véritable probabilisation de l'espace des propriétés.

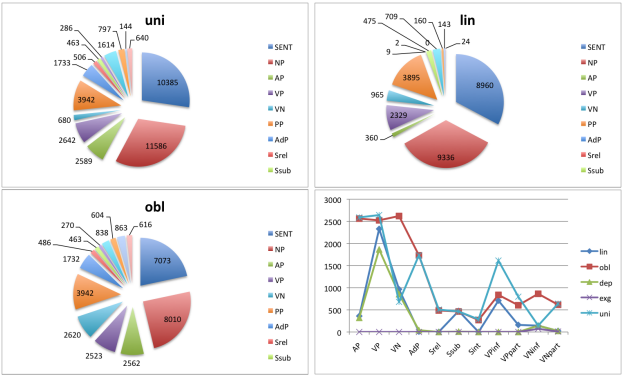

FIGURE 8 – Répartition des propriétés par type

moins denses.

L'étude des caractérisations des principaux syntagmes les plus variables (voir figure 9) fait apparaître trois grands types de répartition : *SENT* et *NP* qui présentent un certain équilibre entre les différentes propriétés ; *PP* et *VP* dont les description n'utilisent pas de propriété d'exigence (cooccurrence obligatoire de catégories) ; *AP* et *AdP* enfin qui n'utilisent quasiment que des propriétés d'unicité et d'obligation.

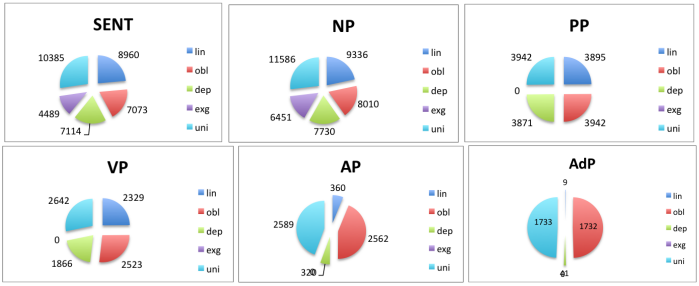

FIGURE 9 – Répartition des propriétés par catégorie

Le niveau le plus fin de description consiste à analyser pour chaque syntagme la répartition des propriétés prises individuellement. Celle-ci n'est en effet pas homogène, ce qui nous permet d'identifier avec précision quels types de propriété jouent un rôle plus important que les autres. Une estimation sur la base de la fréquence constitue un élément indicatif sur la base de laquelle une telle estimation peut être faite. Les tableaux de la figure 10 fournissent ces résultats. Ces schémas présentent en abscisses les indices des propriétés dans la grammaire et en ordonnées

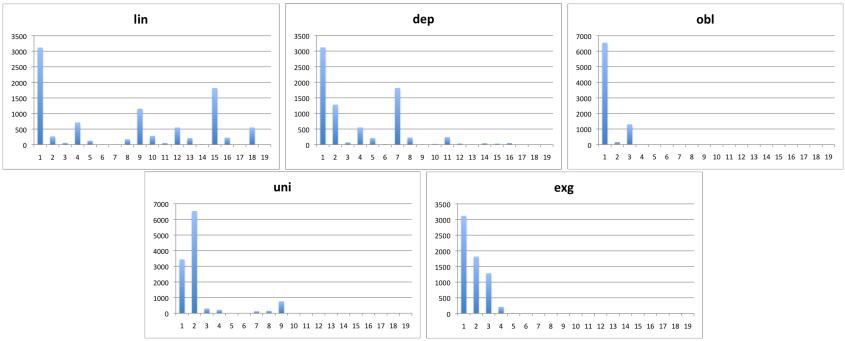

FIGURE 10 – Répartition des propriétés pour le NP

leurs occurrences.

Nous choisissons de retenir comme propriétés à poids fort celles présentant plus de 500 occurrences. Le tableau de la figure 11 récapitule les propriétés fortes du *NP*.

Ce résultat nous offre la possibilité d'acquérir automatiquement une méthode de calcul des poids des propriétés pour chacune des constructions. Cette information, intégrée dans la grammaire, est extrêmement utile tout d'abord en termes de contrôle du processus d'analyse : les propriétés fortes seront traitées comme impérativement satisfaites tandis que les faibles pourront être relâchées. Par ailleurs, du point de vue de la modélisation des processus cognitifs, il s'agit également d'une information très importante, contribuant notamment à l'évaluation de la difficulté de traitement (la violation de contraintes fortes entraînera une difficulté de traitement plus importante).

8 Conclusion

Nous avons décrit dans cet article un processus d'enrichissement du *FTB* permettant de produire un treebank hybride incluant les analyses en *Grammaires de Propriétés*. Le résultat constitue

Type	Indice	Propriété
Linéarité	1	Det $\prec$ N
	4	Det $\prec$ AP
	9	Det $\prec$ PP
	15	N $\prec$ Srel
	18	N $\prec$ NP
Exigence	1	Det $\Rightarrow$ N
	2	PP $\Rightarrow$ N
	3	AP $\Rightarrow$N

Type	Indice	Propriété
Dépendance	1	Det $\rightsquigarrow$ N
	2	AP $\rightsquigarrow$ N
	4	NP $\rightsquigarrow$ N
	7	PP $\rightsquigarrow$ N
Obligation	1	N
	3	Pro
Unicité	1	Det
	2	N

FIGURE 11 – Liste des propriétés *fortes* du NP

une nouvelle ressource pour la description du français à partir de laquelle le développement de nouveaux outils sera possible. Le processus d'enrichissement est totalement automatique, il est donc possible d'envisager la constitution de nouveaux treebanks hybrides. Par ailleurs, ce type treebank inclut une évaluation quantifiée de la grammaticalité grâce à la description en propriétés, ce qui en fait une ressource précieuse notamment pour les études en psycholinguistiques sur corpus.

Références

Abeillé, A., Clément, L. et F., T. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*, Kluwer, Dordrecht.

Blache, P. (2001). *Les Grammaires de Propriétés : Des contraintes pour le traitement automatique des langues naturelles*. Hermès.

Blache, P., Hemforth, B. et Rauzy, S. (2006). Acceptability prediction by means of grammaticality quantification. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 57–64, Sydney, Australia. Association for Computational Linguistics.

Blache, P. et Rauzy, S. (2008). Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks. *In Actes de Traitement Automatique des Langues Naturelles*, pages 290–299, Avignon, France.

Candito, M.-H., Crabbé, B. et Denis, P. (2010). Statistical french dependency parsing : treebank conversion and first results. *In LREC'2010*.

Cerisara, C., Gardent, C. et Anderson, C. (2010). Building and Exploiting a Dependency Treebank for French Radio Broadcast. *In TLT9 – the ninth international workshop on Treebanks and Linguistic Theories*, Tartu, Estonie.

Ide, N. et Suderman, K. (2007). Graf : A graph- based format for linguistic annotations. *In First Linguistic Annotation Workshop*.

Ide, N. et Véronis, J. (1994). MULTEXT : Multilingual text tools and corpora. *In Proceedings of the 15th. International Conference on Computational Linguistics (*Coling 94*)*, volume I, pages 588–592, Kyoto, Japan.

ISO24612 (2008). Language resource management - linguistic annotation framework. *In ISO/TC 37/SC 4 N522/WG 1/CD 24612*.

Kay, P. et Fillmore, C. (1999). Grammatical Constructions and Linguistic Generalizations : the *what's x doing y ?* Construction. *Language*, 75(1):1–33.

Pynte, J., Abeillé, A. et Toussenel, F. (2001). Constituent length and attachment preferences in french. *In AMLAP'2001*.

Rauzy, S. et Blache, P. (2009). Un point sur les outils du lpl pour l'analyse syntaxique du français. *In Actes du workshop ATALA 'Quels analyseurs syntaxiques pour le français ?'*, pages 1–6, Paris, France.

Schluter, N. et van Genabith, J. (2007). Preparing, restructuring and augmenting a french treebank : Lexicalized parsers or coherent treebanks ? *In Proceedings of PACLING 07*, Melbourne, Australia.