Influence de l'étiquetage syntaxique des têtes sur l'analyse en dépendances discontinues du français

Ophélie Lacroix[1]
(1) LINA - Université de Nantes, 2 Rue de la Houssinière, 44322 Nantes Cedex 3
ophelie.lacroix@univ-nantes.fr

RÉSUMÉ

Dans cet article nous souhaitons mettre en évidence l'utilité d'un étiquetage syntaxique appliqué en amont d'une analyse syntaxique en dépendances. Les règles de la grammaire catégorielle de dépendances du français utilisées pour l'analyse gèrent les dépendances discontinues et les relations syntaxiques à longue distance. Une telle méthode d'analyse génère un nombre conséquent de structures de dépendances et emploie un temps d'analyse trop important. Nous voulons alors montrer qu'une méthode locale d'étiquetage peut diminuer l'ampleur de ces difficultés et par la suite aider à résoudre le problème global de désambiguïsation d'analyse en dépendances. Nous adaptons alors une méthode d'étiquetage aux catégories de la grammaire catégorielle de dépendance. Nous obtenons ainsi une pré-sélection des têtes des dépendances permettant de réduire l'ambiguïté de l'analyse et de voir que les résultats locaux d'une telle méthode permettent de trouver des relations distantes de dépendances.

ABSTRACT

On the Effect of Head Tagging on Parsing Discontinuous Dependencies in French

In this paper we want to show the strong impact of syntactic tagging on syntactic dependency parsing. The rules of categorial dependency grammar used to parse French deal with discontinuous dependencies and long distance syntactic relations. Such parsing method produces a substantial number of dependency structures and takes too much parsing time. We want to show that a local tagging method can reduce these problems and help to solve the global problem of dependency parsing disambiguation. Then we adapt a tagging method to types of the categorial dependency grammar. We obtain a dependency-head pre-selection allowing to reduce parsing ambiguity and to see that we can find distant relation of dependencies through local results of such method.

MOTS-CLÉS : Analyse syntaxique en dépendances discontinues, Étiquetage syntaxique.

KEYWORDS: Discontinuous Dependency Parsing, Syntactic Tagging.

1 Introduction

L'analyse syntaxique est une tâche bien connue dans le domaine du traitement automatique du langage naturel, permettant d'obtenir des structures syntaxiques à partir de phrases du langage naturel. On oppose couramment les représentations syntaxiques des structures par consituants et des structures en dépendances. Ici, nous nous intéressons particulièrement à la représentation en dépendances de ces structures (Tesnière, 1959; Mel'cuk, 1988). En utilisant cette représentation, nous souhaitons exprimer correctement les relations syntaxiques existantes entre les mots d'une phrase. Ces relations sont des relations binaires (dépendances) entre un gouverneur g et un subordonné s où le type de dépendance d est la fonction syntaxique existante entre g et s ($g \xrightarrow{d} s$). Une telle dépendance est projective si chaque mot dans l'intervalle [g,s] dépend de g (sinon elle est discontinue). Le type de dépendance d est aussi la dépendance-tête [1] du subordonné s. Notre travail se situe au niveau de l'analyse syntaxique en dépendances pour le français. Or cette langue admet des cas de discontinuité à travers des relations de longue distance comme la coréférence (voir figure 1) ou la comparaison ou des relations locales fréquentes, par exemple de négation ou de clitique. Nous avons choisi une méthode d'analyse guidée par les règles d'une

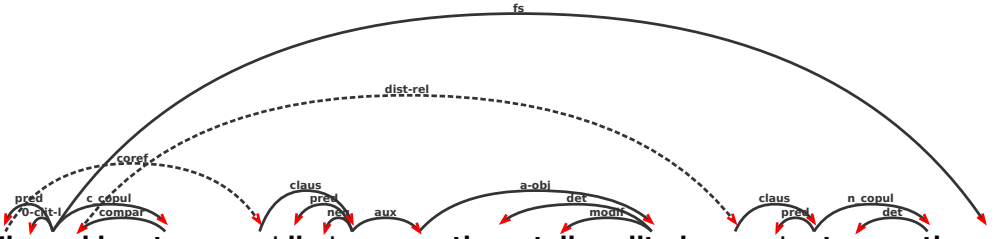


FIGURE 1 – *Structure de dépendances pour la phrase* "Il y a si longtemps qu'il n'a ressenti une telle solitude que c'est un vertige.". *Les dépendances projectives sont représentées par des lignes plaines tandis que les dépendances discontinues sont représentées par des lignes pointillées. Les types des dépendances sont les types utilisés par une grammaire catégorielle de dépendances du français.*

grammaire permettant d'obtenir des structures de dépendances projectives et des structures de dépendances discontinues [2]. Le modèle de grammaire catégorielle de dépendances (Dikovsky, 2004; Béchet *et al.*, 2005; Dekhtyar et Dikovsky, 2008; Dekhtyar *et al.*, 2012) étend la gestion des dépendances aux dépendances discontinues et est donc tout à fait adaptée à la représentation syntaxique en dépendances de phrases du français. Le CDG Lab (Alfared *et al.*, 2011) est un outil, destiné à l'analyse syntaxique avec des grammaires catégorielles de dépendances et au développement de corpus arborés en dépendances. Il propose trois modes d'analyses différents que nous redéfinirons par la suite. Le mode nous intéressant ici est le mode semi-automatique de *sélection des têtes*. Dans ce mode, un utilisateur souhaitant procéder à une analyse syntaxique en

1. La dépendance-tête est le type de la dépendance arrivant sur le subordonné.

2. Une structure de dépendances discontinue est une structure dans laquelle on trouve au moins une dépendance discontinue. Dans ces structures les dépendances peuvent se croiser. Par exemple, les clitiques engendrent des dépendances discontinues dès lors qu'une forme composée verbale est employée, séparant le verbe et son objet cliticisé. La négation produit fréquemment une discontinuité puisqu'elle est communément composé de deux particules, parfois distantes ("Ne ... que"), parfois inversés ("Jamais ... ne"). Par ailleurs, la relation de coréférence (figure 1) est intraphrasale, elle correspond à la co-prédication définie par (Mel'cuk, 1988).

dépendances pourra sélectionner manuellement les dépendances-têtes. Cette sélection des têtes en amont de l'analyse syntaxique améliore grandement la vitesse d'analyse par rapport à une analyse automatique à partir des phrases brutes du français. Nous souhaitons donc remplacer, pour notre travail, cette sélection des têtes manuelles par une sélection automatique. Cette tâche est similaire à celle d'étiquetage grammaticale ou d'étiquetage morphosyntaxique. L'idée d'utiliser un pré-étiquetage pour réduire l'ambiguïté et améliorer une analyse syntaxique en dépendances à déjà été exploitée dans ce cadre (Nasr, 2006; Candito *et al.*, 2010). Ici, nous souhaitons mettre en place un procédé de type *supertagging* (Bangalore et Joshi, 2010b). La sélection des têtes est en fait un étiquetage syntaxique des unités lexicales des phrases du français adapté à la grammaire catégorielle de dépendance du français utilisée pour l'analyse en dépendance. Il ne s'agit donc pas d'apporter des informations grammaticales ou morphosyntaxiques (propres aux unités lexicales) à l'analyseur mais bien d'apporter des informations syntaxiques qui définissent des fonctions binaires entre unités lexicales. La difficulté est alors de trouver les bonnes étiquettes syntaxiques de manière locale, avec des informations locales, bien que la fonction syntaxique auquelle l'étiquette réfère concerne deux unités lexicales potentiellement distantes. Nous procéderons alors dans un premier temps à cet étiquetage syntaxique en utilisant la méthode des CRF [3] adaptée aux types de dépendances de la grammaire catégorielle de dépendances du français. Nous essayons ici d'utiliser une méthode locale pour résoudre un problème global. Il s'agit de la principale difficulté de cette méthode. Nous souhaitons donc voir si elle permettra d'obtenir les bonnes dépendances-têtes. Puis nous exécuterons l'analyse en dépendances sur les phrases ainsi étiquetées pour constater l'effet positif de cet étiquetage sur le temps d'analyse et sur la production (les structures de dépendances sortantes) de l'analyseur. Pour conclure, nous nous questionnerons sur la place de cette méthode dans une analyse totalement autonome. Est-elle suffisante vis-à-vis des résultats obtenus ou peut-elle être associée à d'autres procédés permettant de combler les imperfections de celle-ci ?

Ce travail s'inscrit dans un travail de plus grande envergure qui comprendra un travail de découpage des phrases en unités lexicales, ainsi que leur étiquetage grammaticale, un travail d'étiquetage syntaxique précédant celui d'analyse syntaxique en dépendances, puis finira par un travail concernant le tri des structures de dépendances en sortie de l'analyseur. Ici nous nous intéressons en particulier à l'étiquetage syntaxique. Nous supposons donc avoir en entrée un bon découpage des phrases en unités lexicales composées ainsi qu'un bon étiquetage grammatical de ces unités.

2 Grammaires catégorielles de dépendances

Le modèle de grammaires catégorielles de dépendances est un modèle de grammaires similaire aux grammaires catégorielles classiques (Bar-Hillel *et al.*, 1964) auxquelles est ajoutée la notion de valence polarisée permettant d'introduire les dépendances discontinues. Dans ce modèle, les catégories sont des types de dépendances et permettent de représenter les dépendances projectives tandis que les valences polarisées sont des types de dépendances associées à des polarités duales ($\nearrow\searrow$et $\swarrow\nwarrow$) permettant de représenter les dépendances discontinues (voir (Dekhtyar et Dikovsky, 2008)). Les règles utilisées par cette classe de grammaires sont présentées dans la figure 1. Les structures de dépendances produites à l'aide de ces grammaires sont alors des graphes orientés acycliques.

3. *Conditional Random Fields* en anglais ou champs markovien conditionnels en français.

$\mathbf{L}^1$	$C^{P_1}\left[C\backslash\beta\right]^{P_2} \vdash \left[\beta\right]^{P_1P_2}$
$\mathbf{I}^1$	$C^{P_1}\left[C^*\backslash\beta\right]^{P_2} \vdash \left[C^*\backslash\beta\right]^{P_1P_2}$
Ω^1	$\left[C^*\backslash\beta\right]^{P} \vdash \left[\beta\right]^{P}$
$\mathbf{D}^1$	$\alpha^{P_1(\swarrow C)P(\nwarrow C)P_2} \vdash \alpha^{P_1PP_2}$, si $(\swarrow C)(\nwarrow C)$ satisfait le principe **FA**

TABLE 1 – *Règles gauches des grammaires catégorielles de dépendances. Des règles symétriques sont utilisées dans le cas des dérivations à droite. Les règles **L, I** et Ω permettent d'éliminer les catégories classiques et les catégories itérables (i.e. dérivables infiniement), et de concaténer ou conserver les valences polarisées en une chaîne que l'on appelle* potentiel*. L'élimination des valences dans la dérivation (règle **D**) se fait sur le principe **FA** (First Available) : les valences duales les plus proches dans un potentiel sont éliminées en premier.*

2.1 Données de la grammaire catégorielle de dépendances du français

Pour notre travail, nous utiliserons une grammaire catégorielle de dépendances du français (Dikovsky, 2011). Elle est constituée d'un ensemble conséquent de règles elles-mêmes composées de types de dépendances (les catégories de la grammaire). Ces types de dépendances, 117 au total, représentent un vaste champ de fonctions syntaxiques exprimant les particularités du français. Ces nombreux types de dépendances sont rassemblés en 39 groupes de dépendances selon leurs fonctions syntaxiques. Par exemple, les dépendances de type objet accusatif (*a-obj*), objet datif (*d-obj*), objet génitif (*g-obj*) sont réunies dans le groupe des objets : *OBJ*. Par ailleurs, les types de dépendances peuvent être associés à des dépendances discontinues, on en compte 27. Parmi les types associés aux dépendances discontinues on trouve ceux appartenant aux groupes des clitiques (*CLIT*), des modifieurs (*MODIF*), des réflexifs (*REFLEX*), des coréférences (*COREF*) et appositions (*APPOS*), des éléments de négation (*NEG*), des aggrégations (*AGRR*), etc.

En outre, les règles de la grammaire sont associées à des classes grammaticales. Lors d'une analyse, avec le choix des règles se fait le choix de ces classes grammaticales et des traits morphologiques des unités lexicales établies en fonction des valeurs des traits employés par le Lefff[4] (Sagot, 2010). On dénombre 185 classes grammaticales.

2.2 CDG Lab : Analyseur en dépendances

Le CDG Lab (Alfared *et al.*, 2011) est un outil de travail dédié à l'analyse en dépendances guidée par les règles de grammaires catégorielles de dépendances. L'analyseur en dépendances du CDG Lab propose 3 modes d'analyse différents mais complémentaires :

– *l'analyse autonome* est un mode permettant de lancer l'analyse à partir d'une phrase du français sans indiquer manuellement d'informations complémentaires. La phrase est donc découpée en mots qui sont eux-mêmes réassociés en unités lexicales possibles[5]. L'analyse (basée sur un algorithme CYK modifié) est alors exécutée à partir de ce découpage.
– *l'analyse par sélection des têtes* est un mode semi-automatique. Avant de procéder à l'analyse, l'utilisateur à la possibilité de choisir les bonnes unités lexicales, leurs classes grammaticales et leurs dépendances-têtes. L'analyse peut ensuite être lancée en tenant compte de ces choix.

4. Lexique des formes fléchies du français.
5. Basées sur Lefff (Sagot, 2010).

– *l'analyse par approximation* s'effectue à la suite d'une analyse automatique ou d'une analyse par sélection des têtes. Elle permet d'annoter positivement ou négativement les attributions des classes grammaticales et/ou des types de dépendances. Appliqué autant de fois que nécessaire, ce mode permet de raffiner la production de l'analyse : la(les) structure(s) de dépendances résultante(s).

Le mode qui nous intéresse ici est celui de la sélection des têtes en amont de l'analyse syntaxique en dépendances. Nous savons que choisir manuellement les dépendances-têtes d'une phrase avant analyse permet de réduire l'ambiguïté en faisant converger l'analyse vers un ensemble de solutions plus restreint. Les avantages se remarquent au niveau du temps de calcul de l'analyse et au niveau de la production de l'analyseur, celui-ci produisant moins de structures de dépendances en sortie. Notons que la sélection des têtes peut se faire au niveau des types de dépendances ou des groupes de dépendances. On appellera alors respectivement : **dépendance-tête** ou **groupe-tête**, le type ou le groupe de dépendances sélectionné pour une unité lexicale. Nous souhaitons donc remplacer la sélection manuelle des têtes par une sélection automatique et comprendre l'apport réel de cette tâche avec un algorithme standard d'apprentissage.

2.3 Corpus en dépendances, données grammaticales et syntaxiques

Le corpus que nous utiliserons pour nos expérimentations a été annoté en dépendances semi-automatiquement grâce à l'outil CDG Lab. Il est composé de 2778 structures de dépendances associées à des phrases du français provenant de registres variés et comprenant au total 35203 unités lexicales composées. Les dépendances discontinues représentent 4% du nombre total de dépendances du corpus et elles sont présentes (au moins une fois) dans 41% des structures de dépendances.

Les types de dépendances utilisés dans la représentation de ces structures correspondent aux types de la grammaire catégorielle de dépendances du français. De plus, chaque unité lexicale dans le corpus est annotée correctement par une classe grammaticale. Pour procéder à l'étiquetage syntaxique les données utilisées seront :

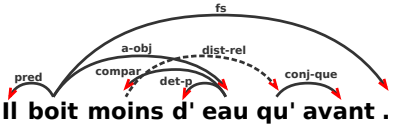
– les unités lexicales composées
– les dépendances-têtes (ou groupes-têtes)
– les classes grammaticales

Le nombre de classes grammaticales étant important, nous décidons de sous-catégoriser ces classes pour arriver à deux formes de sous-classification : les classes grammaticales simples (28 classes) et les classes grammaticales étendues (86 classes). Une classe grammaticale simple indique la classe grammaticale d'une unité lexicale sans autre information tandis qu'une classe grammaticale étendue ajoute des informations (parfois sémantiques) potentiellement utiles syntaxiquement. Les classes grammaticales étendues, plus précises, permettent de mieux cibler les types et groupes de dépendances comme exposé dans la table 2.

Nombre moyen de		types	(max.)	groupes	(max.)
Par classe	simple	13	(43)	7	(18)
grammaticale	étendue	6	(31)	4	(16)

TABLE 2 – *Nombre moyen (et maximum) de types et groupes de dépendances possibles par classe grammaticale simple ou étendue.*

Un exemple, regroupant les données par groupe/type de dépendances et par classe grammaticale simple et étendue, est donné par la figure 2.



Unité lexicale	Dépendance-tête type	groupe	Classe grammaticale simple	étendue
Il	pred	PRED	PN	PN-pers-n
boit	S	SENT	Vt	Vt-fin
moins	compar	COMPAR	Adv	Adv-degr-compar
d'	det-p	DET	Det	Det
eau	a-obj	OBJ	N	N
qu'	dist-rel	REL	Conj	Conj-comp
avant	conj-que	CONJ	Adv	Adv
.	fs	PUNCT	FullStop	FullStop

FIGURE 2 – *Structure de dépendances et tableau rapportant les dépendances-têtes, les groupes-têtes, les classes grammaticales simples et les classes grammaticales étendues de chaque unité lexicale de la phrase* "Il boit moins d'eau qu'avant." *Les classes grammaticales étendues ajoutent des informations en plus de la classe. Par exemple,* Adv-degr-compar *et* Conj-comp *indique un adverbe et une conjonction qui sont tous deux impliqués dans une comparaison. Pour ce travail, nous n'avons pas utilisé les traits de Lefff.*

3 Étiquetage syntaxique

Le problème de l'étiquetage est un problème largement étudié dans le domaine du traitement automatique de la langue naturelle. Les tâches d'étiquetage grammatical ou morphosyntaxique sont les plus répandues mais diffèrent de l'étiquetage syntaxique. Néanmoins les outils restent les mêmes. Parmi les méthodes existantes pour accomplir la tâche d'étiquetage syntaxique on trouvera, les modèles graphiques probabilistes tels que les modèles de Markov cachés (HMM) (Rabiner, 1989), les modèles d'entropie maximale (MEMM) (Ratnaparkhi, 1996) et les champs markoviens conditionnels (CRF) (Sutton et McCallum, 2006; Lafferty *et al.*, 2001). Pour notre travail, nous avons choisi d'utiliser ces derniers car ils permettent de prendre en compte plus d'informations que les HMM et qu'ils sont bien adaptés à l'attribution de séquences d'étiquettes alors que les MEMM sont plus performants pour la classification.

3.1 Logiciel et patrons de traits

Nous avons choisi le logiciel Wapiti (Lavergne *et al.*, 2010) pour entraîner un modèle et étiqueter syntaxiquement notre corpus car il est capable de travailler avec un grand nombre d'étiquettes . Il utilise les CRF pour cet entraînement et attribue donc des poids à des traits choisis. Ces traits peuvent être extraits à partir de patrons de traits définis à l'avance. Le logiciel nous laisse la

possibilité de lui fournir des patrons de traits modifiables que nous avons testés.
Comme indiqué dans la partie 2.3, chaque phrase du corpus est décomposée en unités lexicales elles-mêmes étiquetées grammaticalement (par des classes grammaticales simples ou étendues selon le choix d'expérimentation). Nous disposons donc de ces informations. Nous pouvons choisir une largeur de fenêtre (appliquée autour d'une unité lexicale) pour indiquer si l'on tient compte des unités lexicales et des classes grammaticales précédentes et suivantes lors de l'assignation d'une étiquette syntaxique. Nous constatons qu'une fenêtre de 5 (2 mots avant, 2 mots après) donne de bons résultats, qu'élargir la fenêtre à 7 pour les unités lexicales génère beaucoup de traits pour peu d'améliorations mais qu'élargir la fenêtre à 7 autour des classes grammaticales est beaucoup plus efficace. Il est aussi intéressant d'associer unité lexicale et classe grammaticale dans un même trait. Les premiers patrons de traits choisis sont les suivants :

Unité lexicale courante
Unité lexicale précedente de 1
Unité lexicale précedente de 2
Unité lexicale suivante de 1
Unité lexicale suivante de 2
Classe grammaticale de l'unité lexicale courante
Classe grammaticale de l'unité lexicale précedente de 1
Classe grammaticale de l'unité lexicale précedente de 2
Classe grammaticale de l'unité lexicale précedente de 3
Classe grammaticale de l'unité lexicale suivante de 1
Classe grammaticale de l'unité lexicale suivante de 2
Classe grammaticale de l'unité lexicale suivante de 3
Unité lexicale courante et sa classe grammaticale

Nous testons aussi quelques traits comme l'extraction du suffixe des unités lexicales (testé pour 2, 3 ou 4 lettres) et le fait de savoir si une unité lexicale commence par une majuscule et retenons les suivants :

Suffixe de 3 lettres de l'unité lexicale courante
L'unité lexicale précédente commence-t-elle par une majuscule ?

Notons ici que les traits choisis sont toujours des traits unigrammes, les traits bigrammes générant trop de traits non pertinants. Cependant, pour chaque trait unigramme, la probabilité qu'il apparaîsse avec chacune des étiquettes possibles est calculée lors de l'apprentissage. L'ensemble de ces patrons de traits génère alors plus d'un million [6] de traits pour chaque modèle d'apprentissage et permet d'obtenir de bons résultats d'étiquetage précisés dans la section suivante.

3.2 Expérimentations et Évaluation

Pour procéder à l'étiquetage syntaxique nous avons divisé le corpus (voir section 2.3) en 10 parties égales. Chaque partie est étiquetée selon un modèle entraîné sur les 9 autres parties. L'entraînement se fait sur des données parfaitement étiquetées grammaticalement et syntaxiquement. La possibilité de choisir des données plus ou moins informatives (classe grammaticale simple ou étendue ; dépendance-tête ou groupe-tête) permet de réaliser 4 expérimentations

6. L'ensemble des patrons de traits génère, au pire, plus de 32000 traits unigrammes différents qui associés aux différentes possibilités d'étiquettes (au maximum 117 pour les types) produit jusqu'à 3,7 millions de traits.

différentes. De plus, l'outil Wapiti nous permet d'engendrer les n meilleurs étiquetages pour une séquence donnée. Nous avons donc choisi de produire les 10 meilleurs étiquetages syntaxiques pour chaque phrase d'entrée. Ainsi à chaque expérimentation, nous récupérons 10 séquences d'étiquettes pour chaque phrase du corpus. Ces séquences sont potentiellement assez similaires. Souvent, seulement quelques étiquettes varient d'une séquence à une autre. Pour évaluer la qualité de l'étiquetage syntaxique nous considérons les 1, 2, 5 ou 10 meilleures étiquettes de chaque unité lexicale de chaque phrase du corpus. Les résultats de l'évaluation sont présentés dans la table 3.

Étiquetage des dépendances-têtes

	Classes Grammaticales simples				Classes Grammaticales étendues			
	Pré.	(Moy.)	Rap.	(Moy.)	Pré.	(Moy.)	Rap.	(Moy.)
Top 1	87.8	(70.7)	87.8	(62.9)	91.1	(77.8)	91.1	(70.1)
Top 2	83.4	(66.0)	90.0	(67.5)	86.5	(72.5)	93.2	(74.1)
Top 5	73.0	(56.2)	92.9	(73.4)	75.1	(61.3)	95.5	(79.6)
Top 10	62.9	(46.6)	94.6	(77.2)	63.6	(51.0)	96.6	(82.4)

Étiquetage des groupes-têtes

	Classes Grammaticales simples				Classes Grammaticales étendues			
	Pré.	(Moy.)	Rap.	(Moy.)	Pré.	(Moy.)	Rap.	(Moy.)
Top 1	90.4	(86.5)	90.4	(80.1)	91.6	(89.5)	91.6	(85.6)
Top 2	85.6	(81.0)	92.5	(83.6)	86.8	(83.8)	93.7	(87.9)
Top 5	74.3	(67.7)	95.1	(87.9)	75.0	(71.8)	96.0	(91.2)
Top 10	63.3	(55.2)	96.4	(90.6)	63.4	(57.8)	97.1	(93.1)

TABLE 3 – *Évaluation de l'étiquetage syntaxique produit par Wapiti. D'une part, la précision et le rappel sont calculés globalement sur toutes les étiquettes. La précision est le nombre d'étiquettes correctes sur le nombre d'étiquettes différentes attribuées. Le nombre d'étiquettes différentes attribuées varie selon le top, il peut y en avoir 1, de 1 à 2, de 1 à 5 ou de 1 à 10 (on ne compte pas deux fois la même étiquette). Le rappel est le nombre d'unités lexicales pour lesquelles on a trouvé la bonne étiquette (parmi les 1, 2, 5 ou 10 étiquettes attribuées) sur le nombre d'étiquettes du corpus d'entrée (i.e. le nombre d'unités lexicales). D'autre part, une moyenne de la précision et du rappel sur les types/groupes de dépendances est aussi calculée (entre parenthèses). Dans ce cas, pour chaque type/groupe, la précision est le nombre d'étiquettes correctement attribuées sur le nombre d'étiquettes différentes attribuées pour ce type/groupe. Le rappel pour un type/groupe est le nombre d'unités lexicales y appartenant pour lesquelles on a trouvé la bonne étiquette sur le nombre d'étiquettes de ce type/groupe existantes dans le corpus d'entrée.*

Un permier constat face aux résultats d'étiquetage est de voir l'utilité des informations apportées par les classes grammaticales étendues. De ce côté les résultats sont meilleurs en précision et en rappel. De manière plus approfondie, on peut voir que plus on considère d'étiquettes plus la précision diminue tandis que le rappel augmente. En effet plus on a d'étiquettes différentes pour une unité lexicale plus on a de chance d'avoir la bonne étiquette parmi celles-ci mais on ne sait pas de laquelle il s'agit, on perd donc en précision. En fait, les résultats par étiquette varient grandement. Parmi les 2, 5 ou 10 séquences d'étiquettes pour une même phrase, seulement quelques étiquettes varient. Les étiquettes qui ne changent pas (ou peu) à chaque séquence sont globalement "sûres" et perdent peu en précision (comme les déterminants *DET*, la ponctuation

PUNCT, la négation *NEG* dans le cas des groupes). Celles qui gagnent fortement en rappel sont celles qui sont souvent mal attribuées dans la première séquence mais que l'on finit par trouver dans les suivantes (comme les relations souvent distantes de coréférence *COREF* ou d'apposition *APPOS*). Nous souhaitons voir quel impact a ce gain en rappel sur l'analyse syntaxique en dépendance. Dans la section suivante nous verrons dans quelle mesure l'étiquetage syntaxique réduit le temps d'analyse en dépendance et permet d'obtenir une meilleure structure de dépendances selon les différents critères d'étiquetage que nous avons établis auparavant.

4 Analyse syntaxique en dépendances et évaluation

4.1 Procédure d'analyse et d'évaluation

Pour procéder à l'analyse syntaxique en dépendances nous souhaitons adapter l'outil d'analyse par sélection des têtes du CDG Lab pour assigner automatiquement les étiquettes syntaxiques, trouvées par Wapiti, en tant que dépendances-têtes (ou groupes-têtes). Nous attribuons donc 1, 1 à 2, 1 à 5 ou 1 à 10 types (ou groupes) de dépendances différents à chaque unité lexicale composée selon les résultats des top 1, 2, 5 et 10 de l'étiquetage syntaxique. Nous utilisons ici les meilleurs résultats, c'est à dire ceux trouvés avec les classes grammaticales étendues. L'analyse syntaxique en dépendances guidée par les règles de la grammaire catégorielle de dépendances du français s'éxecute en tenant compte des différentes dépendances-têtes (ou groupes-têtes) possibles. Le CDG Lab est conçu pour produire une liste des structures de dépendances possibles pour chaque analyse [7]. La sélection des têtes permet de réduire l'ambiguïté en contraignant l'analyseur à chercher des structures de dépendances dont les types sont en accord avec cette sélection. Vis-à-vis d'une analyse autonome, ici, le nombre de structures de dépendances en sortie est moindre. Nous observons donc des temps d'analyse également réduits. Les structures de dépendances en sortie de l'analyseur ne sont pas triées. Or nous souhaitons avant tout savoir si parmi les structures de dépendances produites pour une phrase donnée se trouve la bonne structure de dépendances (i.e. la structure de dépendances associée à cette phrase dans le corpus en dépendances de référence, 2.3). L'idée est donc de trier ces structures de la plus proche à la plus éloignée de la structure originale. La plus proche étant celle ayant le plus de dépendances en commun [8] avec la structure de référence. Les différentes étapes de ce traitement sont illustrées dans la figure 3.

Nous nous intéressons alors seulement à la première structure de dépendances de chaque liste (la plus proche de la structure originale). Néanmoins, parfois, il n'existe aucune structure de dépendances produite. Deux raisons sont possibles :

– les dépendances-têtes (ou groupes-têtes) assignées sont en contradiction avec les règles de la grammaire, cela entraîne alors un échec de l'analyse ;
– le temps d'analyse est trop élevé (communément due à la longueur de la phrase), l'analyse s'interrompt donc avant d'aboutir.

Nous souhaitons donc connaître le nombre de structures de dépendances obtenues en sortie. Les résultats de ses expérimentations sont présentés dans la section suivante.

7. En accord avec les dépendances-têtes (ou groupes-têtes) sélectionnées ainsi qu'avec la grammaire.

8. Une dépendance est commune aux deux structures de dépendances si elle possède dans les deux structures le même gouverneur, le même subordonné et le même type de dépendance.

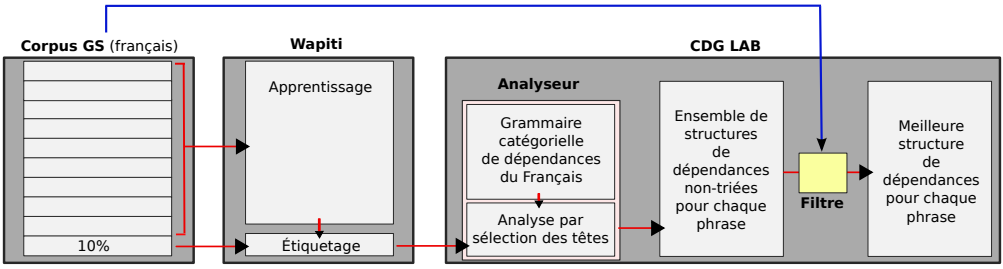


FIGURE 3 – *Schéma expliquatif du traitement complet. Wapiti est utilisé pour procéder à l'apprentissage sur 90% du corpus et à l'étiquetage sur 10%. La partie étiquetée est analysée par l'analyseur du CDG Lab en tenant compte des dépendances-têtes (ou groupes-têtes). On obtient plusieurs structures de dépendances pour chaque phrase qui sont ensuite triées selon leur conformité avec la structure originale du corpus de référence (filtre). Le traitement est opéré sur chaque partie du corpus.*

4.2 Résultats et discussions

Les premiers résultats exposés dans la table 4 présente les taux d'analyses abouties ainsi que le nombre d'unités lexicales par phrase et le temps de calcul.

Nous rapportions dans la section 3.2 qu'en ayant plus de choix de têtes pour chaque unité lexicale nous avions plus de chance d'obtenir la bonne tête parmi ceux-ci. Il en est de même pour les dépendances lorsqu'on laisse entre 1 à 10 choix de têtes pour chaque unité lexicale : ayant plus de chance d'avoir les bonnes dépendances-têtes (ou groupes-têtes) nous avons aussi plus de chance d'obtenir une structure de dépendances proche de la structure de dépendances de référence parmi toutes celles produites. Pour la même raison, on obtient de meilleurs taux d'analyses ayant abouti. Dans le meilleur des cas (sélection de 10 étiquettes), nous avons 2548 analyses sur 2778 (91.7%) ayant abouti dont 2088 ayant trouvé, parmi les structures de dépendances produites, une structure de dépendances entièrement correcte. Les temps d'analyse augmentent relativement à l'ambiguïté (en obtenant plus de structures de dépendances en sortie) ainsi qu'à la longueur des phrases. Effectivement les phrases qui sont analysées avec un choix de dix étiquettes alors qu'elles ne l'étaient pas avec un choix inférieur sont souvent plus longues car plus difficiles à étiqueter correctement et exploitent plus de temps d'analyse. Par ailleurs, on constate que le nombre moyen d'unités lexicales par phrase augmente légèrement dans le cas des analyses ayant abouti quand le choix d'étiquettes est plus large. Ce qui montre que des phrases plutôt longues qui n'ont pas été analysées avec un seul choix d'étiquettes l'ont été avec plus de choix. Cependant un nombre d'étiquettes plus important en entrée augmente l'ambiguïté de l'analyse et donc le temps d'analyse. On obtient donc un peu plus de phrases non-analysées par manque de temps lorsqu'on augmente le choix d'étiquettes.
Lorsque l'on compare les résultats des expérimentations faites avec les dépendances-têtes et les groupes-têtes, on constate plusieurs points intéressants. Les taux d'analyses abouties sont meilleurs lorsqu'on utilise les groupes-têtes car on laisse un plus large choix à l'analyseur (les groupes comprennent parfois plusieurs types de dépendances). Néanmoins on note une différence au niveau du temps d'analyse qui est inférieur lorsqu'on utilise les dépendances-têtes. En effet, l'analyseur converge plus vite. On peut donc voir qu'il y a moins d'analyses n'ayant pas abouti par manque de temps mais que le nombre d'analyses n'ayant pas abouti car étant non-conforme

Analyse autonome

Nb de têtes	Nombre de phrases AA	(%)	NA-C	(%)	NA-T	(%)	UL/phrase AA	NA-C	NA-T	Temps d'analyse AA	NA
0	1150	(41.4)	3	(00.1)	1625	(58.5)	7.2	7.3	17.6	42min24	4h30

Analyse avec sélection des dépendances-têtes

Nb de têtes	Nombre de phrases AA	(%)	NA-C	(%)	NA-T	(%)	UL/phrase AA	NA-C	NA-T	Temps d'analyse AA	NA
1	1805	(65.0)	969	(34.9)	4	(00.1)	11.5	16.5	52.5	3min03	1min35
1 à 2	2054	(73.9)	718	(25.8)	6	(00.2)	11.6	17.7	56.1	4min16	1min53
1 à 5	2335	(84.1)	438	(15.8)	5	(00.2)	12.0	20.0	49.4	6min02	1min29
1 à 10	2505	(90.2)	262	(09.4)	11	(00.4)	12.2	22.5	42.8	8min01	2min23

Analyse avec sélection des groupes-têtes

Nb de têtes	Nombre de phrases AA	(%)	NA-C	(%)	NA-T	(%)	UL/phrase AA	NA-C	NA-T	Temps d'analyse AA	NA
1	1931	(69.5)	832	(29.9)	15	(00.5)	11.5	16.9	45.8	6min41	3min31
1 à 2	2172	(78.2)	586	(21.1)	20	(00.7)	11.6	18.6	45.0	8min52	4min15
1 à 5	2439	(87.8)	302	(10.9)	37	(01.3)	11.8	21.6	43.6	12min05	6min47
1 à 10	2548	(91.7)	179	(06.4)	51	(01.8)	12.0	24.4	41.6	16min43	9min03

TABLE 4 – *Calcul du nombre de phrases dont l'analyse a abouti (AA), du nombre de phrases dont l'analyse n'a pas abouti car elle est non-conforme à la grammaire (NA-C), du nombre de phrases dont l'analyse n'a pas abouti par manque de temps (NA-T). Le temps d'analyse est limité à 10s maximum. Calcul du nombre moyen d'unités lexicales (UL) par phrase dont l'analyse a abouti et dont l'analyse n'a pas abouti (car non-conforme ou par manque de temps). Calcul du temps total d'analyse pour celles ayant abouti et celles n'ayant pas abouti.*

à la grammaire est plus élevé. En fait, l'étiquetage est plus précis donc plus rapide mais conduit plus facilement à une analyse non-conforme à la grammaire s'il y a une ou plusieurs étiquettes fausses. On obtient plus facilement une incohérence vis-à-vis de la grammaire.

D'autre part, nous présentons dans la table 5 en tant que score de précision, les scores d'attachement obtenus sur les analyses abouties. On y trouve le pourcentage d'unités lexicales pour lesquelles le bon gouverneur et la bonne étiquette ont été trouvés (LAS) et le taux d'unités lexicales pour lesquelles le bon gouverneur a été trouvé (UAS). Encore une fois, nous pouvons voir que plus large est le choix d'étiquettes en entrée plus les scores sont meilleurs. Ils atteignent globalement des taux élevés et sont quelques peu meilleurs dans le cas où l'on considère seulement l'exactitude du gouverneur. Lorsqu'on s'intéresse aux dépendances discontinues, on remarque que la précision sur ces dépendances est légèrement moins bonne que sur l'ensemble des dépendances.
La différence de précision entre les analyses ayant reçu des dépendances-têtes ou des groupes-têtes est négligeable sur l'ensemble des dépendances mais moins bonne sur les dépendances discontinues dans le cas de la sélection des groupes-têtes. Cela peut s'expliquer par le fait que le taux de dépendances discontinues est moins élevé parmi les dépendances des analyses abouties dans le cas de la sélection des dépendances-têtes [9]. Les cas difficiles de dépendances discontinues distantes sont écartés du calcul de la précision si la pré-sélection des têtes est non-conforme à la

9. On obtient de 4,3% à 4,6% de dépendances discontinues parmis les dépendances des analyses abouties avec sélection des dépendances-têtes pour 4,8% à 4,9% avec la sélection des groupes-têtes.

grammaire et engendre une mauvaise analyse. Nous avons vu que cette non-conformité est plus facile à atteindre avec la sélection des dépendances-têtes. Les scores sont donc moins bons avec la sélection des groupes-têtes car plus d'analyse aboutissent sans forcement avoir résolu les cas discontinus difficiles.

Analyse autonome

Nb de têtes	Toutes dépendances		Dépendances discontinues	
	LAS	UAS	LAS	UAS
0	98.3	99.0	92.7	93.2

Analyse avec sélection des dépendances-têtes

Nb de têtes	Toutes dépendances		Dépendances discontinues	
	LAS	UAS	LAS	UAS
1	93.7	96.7	92.4	93.7
1 à 2	95.1	97.3	94.3	95.5
1 à 5	96.2	97.8	94.4	95.5
1 à 10	96.4	97.9	94.5	95.4

Analyse avec sélection des groupes-têtes

Nb de têtes	Toutes dépendances		Dépendances discontinues	
	LAS	UAS	LAS	UAS
1	93.9	96.7	88.8	93.3
1 à 2	95.1	97.2	90.0	93.7
1 à 5	96.3	97.9	90.5	93.8
1 à 10	96.7	98.0	91.1	94.3

TABLE 5 – *Évaluation de l'analyse autonome (sans pré-sélection des têtes) et de l'analyse avec pré-sélection des dépendances-têtes et des groupes-têtes. Cette évaluation est réalisée sur la meilleure structure de dépendances produite par l'analyseur (i.e. la plus proche de la structure de dépendances de référence) pour chaque analyse aboutie. Les scores d'attachement LAS et UAS correspondent respectivement au score d'attachement avec dépendances étiquetées* (Labeled Attachment Score) *et au score d'attachement avec dépendances non-étiquetées* (Unlabeled Attachment Score). *Ils sont calculés sur toutes les dépendances d'une part et sur les seules dépendances discontinues d'autre part, en excluant les dépendances liées à des signes de ponctuations dans les deux cas.*

4.3 Travaux reliés

Plusieurs travaux ont déjà mis en évidence l'utilité des méthodes de type *supertagging* sur l'analyse syntaxique (Clark et Curran, 2004; Sarkar, 2010). Les résultats de ces travaux sont difficiles à comparer avec d'autres pour plusieurs raisons. D'une part les fonctions syntaxiques utilisées ici pour l'analyse en dépendances diffèrent et sont plus nombreuses que dans les travaux où les dépendances proviennent des têtes de constituants. De plus les dépendances discontinues ne sont pas toujours prises en compte. D'autres part, l'analyse en dépendances n'est ici pas totalement autonome en s'appuyant sur certains pré-requis. Nous pouvons tout de même tenter de rapprocher ces travaux d'autres tâches de *supertagging* pour l'anglais (Nasr et Rambow, 2004) ou pour l'allemand (Foth *et al.*, 2006). Les travaux les plus proches sont sans doute ceux de (Nasr et Rambow, 2010) obtenant une précision de 85,7% pour l'anglais.

5 Conclusion et travaux à venir

Les résultats de l'analyse syntaxique en dépendances contrainte par la sélection des têtes reflète d'une réelle utilité de cette sélection automatique. Dans un premier temps, la sélection des dépendances-têtes ou des groupes-têtes en amont de l'analyse en dépendances permet de réduire de manière significative le temps d'analyse. De nombreuses phrases, d'une longueur conséquente, ne permettant pas d'aboutir à une analyse autonome peuvent être finalement analysées grâce à la sélection des têtes. Ce facteur est très important pour atteindre des taux de réussite (analyses abouties) intéressant et donc des résultats réellement exploitables. Par ailleurs, en étiquetant syntaxiquement les unités lexicales des phrases de 1 à 10 étiquettes différentes, on obtient un bon score en précision. Il nous indique que parmi les structures de dépendances produites par l'analyseur du CDG Lab, on obtient très souvent la bonne structure de dépendances pour une phrase donnée.

Cependant, nous supposons ici avoir un bon découpage des phrases en unités lexicales et un bon étiquetage en entrée ainsi qu'un tri des structures de dépendances en sortie qui s'appuie sur la structure de dépendances de référence. Dans l'idée de mettre en place un analyseur totalement autonome, nous souhaitons, par la suite, faire de ces étapes des tâches automatiques. Nous avons donc l'intention d'ajouter une étape de découpage des unités lexicales et d'étiquetage grammatical de ces unités en amont de l'étiquetage syntaxique présenté dans cet article. Puis nous appliquerons une méthode de tri automatique des structures de dépendances en sortie de l'analyseur permettant de trouver la structure de dépendances la plus proche de la structure de référence sans s'y être référé.

En outre, notons que le taux d'analyse non abouties car l'étiquetage était non-conforme avec la grammaire varie de 6 à 35% du meilleur au pire des cas. Un mauvais étiquetage local peut être la cause de cette non-conformité. Cependant le score général d'étiquetage étant bon (le meilleur est de 97.1 en rappel pour 10 choix d'étiquettes), il est évident que la majorité des étiquettes pour une phrase donnée sont correctes et permettraient d'obtenir une (ou plusieurs) sous-structure(s) de dépendances correcte(s) pour cette phrase. L'évolution de l'analyseur du CDG Lab ira dans ce sens : permettre à l'analyseur de produire des structures de dépendances partielles lorsque la sélection des têtes n'est pas totalement conforme avec la grammaire. La solution partielle pourra ensuite être complétée en appliquant une analyse par approximation. Le nombre de structures de dépendances analysées augmentera et cela permettra d'obtenir de meilleurs taux d'analyses abouties.

Références

ALFARED, R., BÉCHET, D. et DIKOVSKY, A. (2011). "CDG Lab" : a Toolbox for Dependency Grammars and Dependency Treebanks Development. *In Proceedings of DEPLING 2011*, pages 272–281.

BANGALORE, S. et JOSHI, A., éditeurs (2010a). *Complexity of Lexical Descriptions and its Relevance to Natural Language Processing : A Supertagging Approach*. MIT Press.

BANGALORE, S. et JOSHI, A. K. (2010b). *Supertagging : Using Complex Lexical Descriptions in Natural Language Processing*. Mit Press.

BAR-HILLEL, Y., GAIFMAN, C. et SHAMIR, E. (1964). On Categorial and Phrase Structure Grammars. *In Language and information*, pages 99–115. Addison-Wesley.

BÉCHET, D., DIKOVSKY, A. et FORET, A. (2005). Dependency structure grammar. *In Proceedings of LACL 2005*, pages 18–34.

CANDITO, M., CRABBÉ, B. et DENIS, P. (2010). Statistical french dependency parsing : treebank conversion and first results. *In Proceedings of LREC 2010*, pages 1840–1847.

CLARK, S. et CURRAN, J. R. (2004). The importance of supertagging for wide-coverage ccg parsing. *In Proceedings of COLING 2004*, pages 282–288.

DEKHTYAR, M. et DIKOVSKY, A. (2004). Categorial dependency grammars. *In Proceedings of Intern. Conf. on Categorial Grammars*, pages 76–91.

DEKHTYAR, M. et DIKOVSKY, A. (2008). Generalized categorial dependency grammars. *In Trakhtenbrot/Festschrift*, LNCS 4800, pages 230–255. Springer.

DEKHTYAR, M., DIKOVSKY, A. et KARLOV, B. (2012). Iterated dependencies and kleene iteration. *In Formal Grammar 2010/2011*, LNCS 7395, pages 66–81.

DIKOVSKY, A. (2004). Dependencies as categories. *In Proceedings of COLING 2004 Workshop, "Recent Advances in Dependency Grammars"*, pages 90–97.

DIKOVSKY, A. (2011). Categorial dependency grammars : from theory to large scale grammars. *In DEPLING 2011*.

FOTH, K., BY, T. et MENZEL, W. (2006). Guiding a constraint dependency parser with supertags. *In Proceedings of COLING 2006*, pages 289–296.

LAFFERTY, J., MCCALLUM, A. et PEREIRA, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of ICML 2001*.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical very large scale CRFs. *In ACL 2010*.

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

NASR, A. (2006). Grammaires de dépendances génératives probabilistes. modèle théorique et application à un corpus arboré du français. *Traitement Automatique des Langues*, 46(1):115–153.

NASR, A. et RAMBOW, O. (2004). Supertagging and full parsing. *In Proceedings of TAG+7*.

NASR, A. et RAMBOW, O. (2010). Non-lexical chart parsing for tag. *In* (Bangalore et Joshi, 2010a).

RABINER, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *In Proceedings of IEEE 1989*.

RATNAPARKHI, A. (1996). A maximum entropy model for part-of-speech tagging. *In Proceedings of EMNLP 1996*.

SAGOT, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *In Proceedings of LREC 2010*.

SARKAR, A. (2010). Combining supertagging and lexicalized tree-adjoining grammar parsing. *In* (Bangalore et Joshi, 2010a).

SUTTON, C. et MCCALLUM, A. (2006). An introduction to conditional random fields for relational learning. *In Introduction to Statistical Relational Learning*. MIT Press.

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