# Étude des dépendances syntaxiques non projectives en français

Guy Perrier

# Étude des dépendances syntaxiques non projectives en français

**Guy Perrier***

* *Université de Lorraine – LORIA – Campus Scientifique – BP 239 - 54506 Vandœuvre-lès-Nancy cedex - France*

guy.perrier@loria.fr

RÉSUMÉ. *Cet article présente les résultats d'une étude mathématique et linguistique des dépendances syntaxiques non projectives dans des corpus du français annotés selon deux formats de syntaxe en dépendance :* Universal Dependencies (UD) *et* Surface Syntactic Universal Dependencies (SUD). *Cette étude met en évidence le caractère très local des configurations de croisement de deux dépendances qui est une des façons de caractériser la non-projectivité. Elle met aussi en évidence quatre sources linguistiques principales de la non-projectivité : la montée de clitiques, l'extraction profonde, la minimisation de la longueur des dépendances et les couples de mots dépendants distants.*

MOTS-CLÉS : *dépendances non projectives, syntaxe en dépendances, détection de motifs dans un graphe.*

TITLE. *Study of non-projective dependencies in French*

ABSTRACT. *This paper presents the results of a mathematical and linguistic study of non-projective syntactic dependencies in French corpora annotated according to two dependency syntax formats:* Universal Dependencies (UD) *and* Surface Syntactic Universal Dependencies (SUD). *This study highlights the very local character of the configurations of two crossing dependencies, which is one of the ways to characterize non-projectivity. It also highlights four main linguistic sources of non-projectivity : clitic climbing, deep extraction, dependency length minimization and pairs of distant dependent words .*

KEYWORDS: *non projective dependencies, dependency syntax, graph pattern matching.*

## 1. Introduction

Les dépendances non projectives posent problème tant aux linguistes qu'aux informaticiens. Elles compliquent pour les premiers la tâche de conversion des structures en dépendances en structures syntagmatiques (Gaifman, 1965), car la non-projectivité s'oppose à la continuité des syntagmes. La discontinuité dans les structures linguistiques non projectives est aussi un obstacle à leur compréhension, si bien que les dépendances non projectives sont aussi intéressantes à étudier pour les psycholinguistes. Pour les informaticiens, elle complique beaucoup les algorithmes d'analyse syntaxique, même si certains ont développé des méthodes qui permettent de réduire cette complexité (Tapanainen et Jarvinen, 1997 ; McDonald *et al.*, 2005 ; Nivre, 2009 ; Gómez-Rodríguez *et al.*, 2014 ; Straka *et al.*, 2015).

Pour traiter les dépendances non projectives sous ces différents aspects, il est utile de bien les connaître et donc de les étudier plus précisément. Or, s'il existe beaucoup de travaux sur l'analyse syntaxique des arbres de dépendances non projectifs, il existe peu d'études de ces arbres dans les corpus réels tant d'un point de vue mathématique que linguistique. Quelques études ont été menées pour des langues particulières, comme le tchèque (Hajicová *et al.*, 2004), le grec ancien (Mambrini et Passarotti, 2013) et le serbe (Miletic et Urieli, 2017) notamment. La principale étude mathématique générale, à notre connaissance, est la thèse de Havelka (2007) qui exhibe différentes caractérisations des arbres de dépendances non projectifs, pour en déduire des algorithmes de recherche de ceux-ci dans un corpus arboré, ce qui lui permet de comparer la fréquence d'apparition des dépendances non projectives entre 19 langues. Parmi ces langues, il n'y a pas le français.

Pour le français, il est à mentionner une étude de Botalla (2014), qui aborde la non-projectivité à travers l'analyse du flux des dépendances dans un treebank, et une autre de Béchet et Lacroix (2015) sur le corpus, *CDGFr*. Ce dernier ayant été annoté de façon non standard, il est difficile d'établir une comparaison avec d'autres travaux.

Dans cet article, nous présentons une étude que nous avons menée sur les dépendances non projectives en français à partir de corpus annotés. Nous avons effectué cette étude en deux temps. Dans un premier temps, nous avons considéré uniquement l'aspect topologique des annotations, c'est-à-dire la structure formée par les dépendances non projectives, indépendamment de leurs étiquettes linguistiques. Le but était de mesurer le plus précisément possible le caractère local et complexe de la non-projectivité. Parmi toutes les façons de caractériser la non-projectivité, nous avons privilégié le croisement de dépendances. Nous avons pu exhiber un nombre limité de motifs finis formés par les couples de dépendances qui se croisent.

Ces motifs ont constitué le point de départ du second temps, l'étude linguistique, où nous avons alors pris en compte les étiquettes linguistiques des dépendances. Nous avons exploré systématiquement les corpus à la recherche de toutes les occurrences des motifs mis en évidence par l'étude topologique. Nous avons utilisé pour cela l'ou-

til GREW-MATCH[1] qui procède par appariement de graphes. C'est une composante d'un outil plus vaste, GREW[2] utilisé pour la transformation d'annotations et fondé sur la réécriture de graphes à l'aide de règles (Bonfante *et al.*, 2018). Partant de motifs purement structurels, GREW-MATCH permet par l'exploration des corpus de les enrichir linguistiquement pas à pas pour faire apparaître les principaux phénomènes linguistiques responsables sur les corpus donnés de non-projectivité.

Cette étude présente un intérêt tant pour l'analyse syntaxique en dépendances que pour l'annotation de corpus et pour l'analyse linguistique de corpus. Tout d'abord, on sait que la non-projectivité complique beaucoup la tâche d'analyse syntaxique. Même si des progrès importants ont été réalisés ces dernières années, les meilleurs analyseurs sont encore loin de la perfection : Kuhlmann et Nivre (2010) n'analysent correctement que la moitié des dépendances non projectives présentes dans des corpus de l'anglais et de l'allemand et la proportion monte aux deux tiers pour le tchèque.

L'étude topologique vise à obtenir des résultats quantitatifs en termes de fréquence, de localité et de complexité des dépendances non projectives qui permettront d'adapter les algorithmes d'analyse afin de concilier efficacité et expressivité. Cela peut se faire tant dans l'approche fondée sur la recherche d'arbres à portée maximale dans un graphe en suivant les travaux de Corro *et al.* (2016) que dans l'approche fondée sur des systèmes de transition comme le montrent Kuhlmann et Nivre (2006). Pour ce qui est de cette dernière approche, la mise en évidence des phénomènes linguistiques sources de non-projectivité, qui est visée dans la partie linguistique de notre étude, permettra de les confronter avec les règles de transition des systèmes utilisés pour étudier dans quelle mesure ces règles peuvent les prendre en compte. D'ailleurs, Kuhlmann et Nivre (2010) insistent sur cet aspect dans la conclusion de leur article : *« Although the experiments presented in this article have already revealed significant differences both between languages and between techniques, it would be interesting to look in more detail at the different linguistic constructions that give rise to non-projective dependencies.(Bien que les expériences présentées dans cet article aient déjà révélé des différences significatives entre les langues et entre les techniques, il serait intéressant d'examiner plus en détail les différentes constructions linguistiques qui donnent lieu à des dépendances non projectives) ».*

Maintenant, plutôt que de chercher à améliorer l'analyse syntaxique, il peut être plus facile et plus efficace de chercher après coup à corriger les dépendances non projectives mal annotées. Depuis plusieurs années, nous avons développé une expertise dans ce domaine en utilisant l'outil GREW-MATCH pour repérer les constructions erronées dans un treebank et l'outil GREW pour corriger les erreurs quand elles sont systématiques (Guillaume *et al.*, 2019). L'étude que nous présentons ici va permettre de dégager des motifs associés aux phénomènes linguistiques responsables de non-projectivité. En appliquant ces motifs à des corpus du français, il est possible tout d'abord de détecter les constructions faussement non projectives. Si le format d'anno-

1. `http://match.grew.fr/`
2. `http://match.grew.fr/`

tation n'est pas celui utilisé dans notre étude, cela exigera quelques adaptations. Pour les constructions faussement projectives, les choses sont un peu plus compliquées car il faudra recenser les erreurs possibles pour les traduire sous forme de motifs.

Une autre application concerne la transformation d'un corpus annoté syntaxiquement en constituants en un corpus annoté en dépendances. Il existe un algorithme classique qui permet de le faire mais le résultat est nécessairement un ensemble d'arbres projectifs. Pour faire apparaître la non-projectivité, il est nécessaire de compléter l'application de cet algorithme par une transformation de certaines dépendances projectives en dépendances non projectives. Candito *et al.* (2009) l'ont fait manuellement pour le FRENCH TREEBANK. Notre étude serait un point de départ à la détermination de règles de réécriture qui permettraient de le faire automatiquement.

Enfin, notre travail peut aider à une étude linguistique plus poussée de la non-projectivité. Il serait tout d'abord intéressant de confronter les phénomènes linguistiques sources de la non-projectivité à différentes théories linguistiques pour étudier dans quelle mesure ces dernières sont capables de fournir une explication. Et puis la méthode que nous présentons est suffisamment simple pour être utilisée par des linguistes qui voudraient faire une étude approfondie de la non-projectivité sur d'autres corpus que ceux que nous avons choisis.

Pour notre étude, nous avons choisi trois treebanks :

– UD_FRENCH-GSD[3] (Guillaume *et al.*, 2019) dont les données proviennent de l'Universal Dependency Treebank v2.0 de Google (McDonald *et al.*, 2013) ; il comprend 16 341 phrases d'origines très diverses (dépêches de presse, blogs, avis de consommateurs. . .) ; à partir de 2015, l'annotation du corpus a été convertie dans le format UD, sur lequel est fondé le projet Universal Dependencies[4] ; ce projet a pour but de créer un schéma d'annotation syntaxique unique qui puisse être utilisé pour un maximum de langues différentes (Nivre *et al.*, 2016) ; c'est pour cette raison que le format UD considère les relations syntaxiques comme des relations directes entre mots lexicaux, et les mots fonctionnels comme des marqueurs des mots lexicaux ;

– SUD_FRENCH-GSD [5] qui résulte d'une conversion du corpus précédent dans un nouveau format, le format SUD (Gerdes *et al.*, 2018 ; Gerdes *et al.*, 2019). SUD est une alternative à UD qui utilise de façon plus classique des critères distributionnels pour définir les relations syntaxiques, si bien que les têtes des relations sont plutôt les mots fonctionnels que les mots lexicaux ;

– UD_FRENCH-SEQUOIA[6] est issu du corpus SEQUOIA qui a d'abord été annoté en syntagmes selon le schéma du FRENCH TREEBANK (Abeillé *et al.*, 2019), puis converti en dépendances (Candito et Seddah, 2012b) ; cette annotation en dépendances a été enfin convertie dans le format UD (Guillaume *et al.*, 2019) ; le corpus comprend

3. https://github.com/UniversalDependencies/UD_French-GSD
4. http://universaldependencies.org
5. https://github.com/surfacesyntacticud/SUD_French-GSD
6. https://github.com/UniversalDependencies/UD_French-Sequoia

3 099 phrases de quatre origines différentes : l'agence européenne du médicament, Europarl, le journal régional l'*Est Républicain* et Wikipedia Fr.

Notre étude a porté sur la version 2.7 de ces trois treebanks. L'intérêt d'avoir un même corpus annoté selon deux formats différents, UD_FRENCH-GSD selon UD et SUD_FRENCH-GSD annoté selon SUD, est de pouvoir étudier dans quelle mesure le format d'annotation rend compte de cette propriété de non-projectivité.

L'intérêt d'avoir deux corpus différents, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, annotés dans un même format, UD, est lui de pouvoir étudier dans quelle mesure les dépendances non projectives dépendent du choix du corpus.

Le plan de l'article est le suivant :

– dans la section 2, nous présenterons les formats d'annotation syntaxique UD et SUD en mettant en évidence leurs différences ;

– dans la section 3, nous nous intéresserons à la topologie des configurations formées par les couples de dépendances qui se croisent, mettant en évidence que ces configurations ont un caractère très local ;

– enfin dans la section 4, nous montrerons que principalement quatre phénomènes linguistiques sont principalement source de dépendances non projectives dans les corpus étudiés : la montée de clitiques, l'extraction profonde, la minimisation de la longueur des dépendances et les couples de mots dépendants distants.

## 2. Les formats d'annotation syntaxique UD et SUD

Comme nous le verrons par la suite, le format d'annotation syntaxique joue un rôle important dans l'existence ou non de dépendances non projectives. C'est pourquoi il est nécessaire d'avoir un aperçu des deux formats utilisés dans les trois corpus étudiés.

### 2.1. *Le format* UD

La définition du format UD (Nivre *et al.*, 2016)[7] a été guidée par le souci de son universalité, c'est-à-dire qu'il puisse être utilisé pour toutes les langues. C'est pourquoi il est guidé par la sémantique. Les têtes des syntagmes sont les mots lexicaux, les mots fonctionnels étant rattachés à ceux-ci comme marqueurs. Les relations de dépendances syntaxiques sont donc des relations entre mots lexicaux.

Une caractéristique de UD, qui n'était pas requise par le souci d'universalité, est que les types des relations renvoient non seulement aux fonctions syntaxiques des mots, mais aussi à leurs parties du discours. Ainsi, la relation entre un modificateur d'un nom et ce nom peut être étiquetée `acl`, `advmod`, `amod`, `nmod`, selon que le modificateur est une proposition, un adverbe, un adjectif ou un nom.

7. `https://universaldependencies.org/guidelines.html`

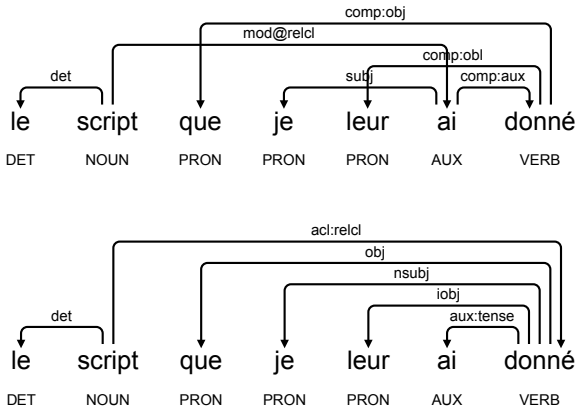


**Figure 1.** *Annotation de la même expression dans* SUD *au-dessus et* UD *au-dessous*

Le schéma du bas de la figure 1 présente un exemple significatif d'une expression annotée dans UD. L'annotation de la phrase fait apparaître un arbre de dépendances très plat avec parfois un nombre important de dépendances se rattachant à un même mot lexical. Ainsi, le mot *donné* a quatre dépendants qui ne sont pas du tout reliés les uns aux autres ; ils sont simplement ordonnés.

### 2.2. *Le format* SUD

Le format SUD (Gerdes *et al.*, 2018 ; Gerdes *et al.*, 2019)[8] se présente comme une alternative au format UD fondée sur l'approche distributionnelle plus classique des relations syntaxiques (Bloomfield, 1933 ; Mel'cuk *et al.*, 1988 ; Kahane et Gerdes, 2020). Dans cette approche, les mots fonctionnels constituent les têtes des syntagmes dans la mesure où ils déterminent leur distribution. Dans SUD, il s'agit des prépositions, des conjonctions de subordination et des auxiliaires. Les conjonctions de coordination et les déterminants, qui jouent un rôle plus discutable dans la distribution des syntagmes qu'ils introduisent, ne sont pas leur tête dans SUD.

Par ailleurs, comme la partie du discours des dépendants n'est pas déterminante dans la distribution des relations, les étiquettes de ces dernières ne la prennent pas en compte et elles ne considèrent que les fonctions syntaxiques.

Enfin, les relations sont organisées en une stricte taxonomie. Un moyen de créer une sous-relation d'une autre est d'ajouter une extension à son nom précédée d'un

8. `https://surfacesyntacticud.github.io/`

deux-points. Ainsi, `comp` représente la fonction argument syntaxique complément en général et `comp:aux` représente la fonction argument d'un auxiliaire.

Le schéma du haut de la figure 1 reprend la même phrase que pour illustrer UD et montre son annotation dans SUD. La différence avec UD est flagrante : les mots fonctionnels sont structurés les uns par rapport aux autres, ce qui augmente la profondeur des arbres de dépendances. Comme l'exemple le montre aussi, cela augmente aussi la possibilité d'avoir des dépendances non projectives.

## 3. Étude topologique des dépendances non projectives en français

Dans cette section, nous nous intéressons aux configurations structurelles formées par les dépendances non projectives indépendamment de leur typage linguistique, effectué en général par l'association des dépendances à des étiquettes représentant des fonctions syntaxiques (sujet, objet. . .). Dans toute la suite de l'étude, nous ignorerons les dépendances qui ciblent des signes de ponctuation, dans la mesure où les guides d'annotation de UD et de SUD sont trop imprécis sur comment choisir leurs gouverneurs, d'où beaucoup d'incohérences dans les corpus existants.

### 3.1. *Caractérisation mathématique de la non-projectivité*

Les structures représentant la syntaxe des phrases considérées dans cet article sont des arbres de dépendances totalement ordonnés. Un *arbre de dépendances totalement ordonné* est un arbre enraciné[9] dont les nœuds sont totalement ordonnés. Par la suite, les arbres de dépendances que nous utiliserons sont toujours totalement ordonnés et nous ne le mentionnerons pas à chaque fois. Les nœuds représentent les mots de la phrase [10]. La relation père-fils dans l'arbre représente la dépendance syntaxique entre les mots ; elle est notée $\rightarrow$. Sa clôture transitive est notée $\rightarrow^{+}$et sa clôture transitive et réflexive $\rightarrow^{*}$. Une instance $\texttt{H} \rightarrow \texttt{D}$ de la relation $\rightarrow$ est appelée *une dépendance*.

La relation d'ordre total entre les nœuds est notée $\leq$ quand elle est entendue au sens large et $<$ quand elle est entendue au sens strict. Elle représente l'ordre des mots dans la phrase.

Historiquement, la distinction entre projectivité et non-projectivité d'un arbre de dépendance totalement ordonné a été mise en évidence par Harper et Hays (1959) et

9. Un arbre enraciné est un graphe connexe acyclique dont on a choisi un nœud particulier comme racine.

10. Dire que les nœuds représentent les mots de la phrase est une simplification car en général, les tokens résultant du découpage de la phrase ne coïncident pas toujours avec les mots de la phrase au sens linguistique. Certains mots peuvent être formés de plusieurs tokens, ce qui est indiqué dans l'arbre syntaxique par des relations de dépendances spécifiques. Ces relations n'étant jamais responsables de non-projectivité, on peut considérer sans perte de généralité que la notion de token coïncide avec celle de mot.

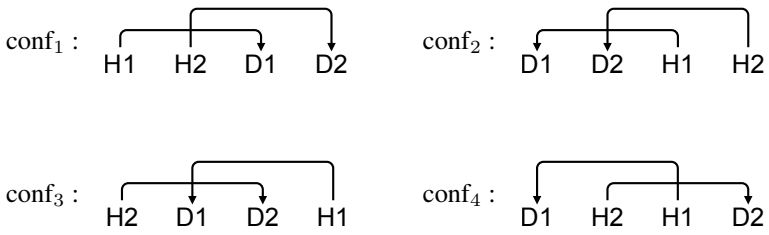


**Figure 2.** *Les quatre configurations possibles de croisement de deux dépendances*

Lecerf et Ihm (1960). Elle a été étudiée par Marcus (1965) qui a démontré l'équivalence entre trois caractérisations d'un arbre de dépendances projectif :

– chacun de ses nœuds `N` est projectif, c'est-à-dire que l'ensemble des nœuds `M` tels que `N` $\rightarrow^*$`M` forme un segment continu selon l'ordre des mots de la phrase (Fitialov, 1962) ; le nœud *donné* de l'arbre SUD de la figure 1 est non projectif car sa projection { *que, leur, donné* } est discontinue ;

– chacune de leurs dépendances `H` $\rightarrow$ `D` est projective, c'est-à-dire que l'ensemble des nœuds situés entre `H` et `D` est inclus dans la projection de `H` (Harper et Hays, 1959) ; l'arbre SUD de la figure 1 comporte deux dépendances non projectives : (`donné -[comp:obj]`$\rightarrow$ `que`) et (`donné -[comp:obl]`$\rightarrow$ `leur`) ;

– les dépendances de l'arbre complété ne se croisent jamais selon l'ordre de la phrase ; l'arbre complété est l'arbre de racine `R` auquel a été ajoutée une dépendance `R'` $\rightarrow$ `R` issue d'une racine fictive `R'` ajoutée à gauche de la phrase [11] ; l'arbre SUD de la figure 1 comporte trois croisements de dépendances.

L'inconvénient de la première caractérisation est qu'elle nécessite de calculer la projection complète des nœuds. La seconde est plus locale puisqu'elle demande seulement à vérifier que tous les nœuds situés entre chaque nœud et chacun de ses dépendants sont dans la projection de ce nœud. La dernière est encore plus facilement calculable car il s'agit de vérifier que chaque paire de dépendances ne donne pas lieu à un croisement de celles-ci, et c'est cette caractérisation que nous avons privilégiée dans notre étude. La figure 2 traduit graphiquement les quatre configurations possibles de croisement de deux dépendances $\mathtt{H}_1 \rightarrow \mathtt{D}_1$ et $\mathtt{H}_2 \rightarrow \mathtt{D}_2$. Les trois croisements de l'arbre SUD de la figure 1 illustrent les configurations 2 et 3.

11. On pourrait tout aussi bien ajouter `R'` à droite de la phrase et Kahane et Gerdes (2020) évitent cet arbitraire en plaçant les mots de la phrase sur un cercle et en ajoutant le nœud `R'` entre le premier et le dernier mot de la phrase.

### 3.2. *Mesure du degré de non-projectivité d'un corpus*

Dans un souci de concevoir des algorithmes d'analyse syntaxique qui respectent un équilibre entre efficacité et pouvoir d'expression, il est utile de connaître la proportion de nœuds non projectifs ou de dépendances non projectives dans un corpus.

On peut aller plus loin en introduisant des paramètres qui mesurent le degré de complexité des nœuds non projectifs et des dépendances non projectives, de telle façon que si ce degré ne dépasse pas une certaine borne, on puisse concevoir des algorithmes d'analyse relativement efficaces. C'est ce qu'ont cherché à faire Kuhlmann et Nivre (2006) et Corro *et al.* (2016). Nous allons reprendre certains de ces paramètres en les appliquant à nos corpus.

Un premier paramètre est le nombre de trous maximal des projections des nœuds d'un arbre de dépendances.

#### 3.2.1. *Nombre maximal de trous dans les projections des nœuds d'un arbre de dépendances*

La notion de trou a été introduite par Holan *et al.* (2000) pour caractériser la complexité des arbres de dépendances non projectifs. Donnons-en une définition formelle.

**Définition 1.** *Un trou dans la projection d'un nœud* `N` *est un ensemble de mots consécutifs qui n'appartiennent pas à la projection de* `N` *et qui est borné à droite et à gauche par deux éléments de cette projection.*

De cette définition, découle immédiatement que tout nœud non projectif est caractérisé par la présence d'un ou plusieurs trous dans sa projection. Ainsi, dans l'annotation SUD de la figure 1, le nœud non projectif *donné* a sa projection [*que, leur, donné*] qui comporte deux trous {*je*} et {*ai*}. Dans l'annotation UD, la projection [*que, je, leur, ai, donné*] du nœud projectif *donné* ne comporte pas de trou.

Nous avons conçu un programme Python qui, pour un corpus donné, calcule le nombre de trous pour la projection de chaque nœud des arbres syntaxiques et, pour chaque arbre, détermine le nombre de trous maximal. Les nœuds projectifs sont ceux dont la projection ne comporte pas de trou et les arbres projectifs sont ceux pour lesquels le maximum du nombre de trous est de 0.

Sur nos trois corpus, nous obtenons les résultats du tableau 1. La première constatation est que selon le format d'annotation, les résultats sont très différents. Les textes des corpus SUD_FRENCH-GSD et UD_FRENCH-GSD sont les mêmes, découpés de la même façon en tokens. Seul diffère le format d'annotation, SUD pour le premier et UD pour le second. Dans SUD_FRENCH-GSD, il y a deux fois et demie plus de nœuds non projectifs que dans UD_FRENCH-GSD. L'explication tient au fait que les arbres de dépendances UD sont beaucoup plus plats que les arbres SUD. Les têtes des syntagmes sont toujours des mots lexicaux et les mots grammaticaux se rattachent directement à ces têtes. Dans SUD au contraire, les têtes des syntagmes sont les mots fonctionnels quand ils sont présents (auxiliaires, prépositions, conjonctions de subor-

| corpus | SUD-GSD | UD-GSD | UD-SEQUOIA | total |
|---|---|---|---|---|
| nb. de nœuds | 356 393 | 356 393 | 62 706 | 775 492 |
| nb. de nœuds non projectifs | 1729 (0,49 %) | 694 (0,19 %) | 74 (0,12 %) | 2 497 (0,32 %) |
| nœuds à 0 trou | 354 664 | 355 699 | 62 632 | 772 995 |
| nœuds à 1 trou | 1 708 | 692 | 73 | 2 473 |
| nœuds à 2 trous | 21 | 2 | 1 | 24 |
| nb. d'arbres syntaxiques | 16 341 | 16 341 | 3 099 | 35 781 |
| nb. d'arbres non projectifs | 1 392 (8,52 %) | 653 (2,62 %) | 66 (2,13 %) | 2 111 (5,90 %) |
| arbres à 0 trou au maximum | 14 949 | 15 688 | 3 033 | 33 670 |
| arbres à 1 trou au maximum | 1 373 | 651 | 65 | 2 089 |
| arbres à 2 trous au maximum | 19 | 1 | 1 | 22 |

**Tableau 1.** *Statistiques sur le nombre de trous dans les projections des nœuds*

dination), et ces mots peuvent être dépendants les uns des autres d'où une structure plus en profondeur des arbres, comme le montre l'exemple de la figure 1.

La seconde constatation est que le maximum du nombre de trous par projection est très bas puisqu'il est de 2 pour les trois corpus considérés. Encore plus intéressant, considérons la répartition des nœuds des arbres d'un corpus selon le nombre de trous par projection. Pour les trois corpus considérés ensemble, on obtient la répartition [772 995, 2 473,24] correspondant à 0 trou, 1 trou, 2 trous. Si on considère les arbres et non plus les nœuds, la répartition est [33 670, 2 089, 22]. La conséquence est que pour avoir des algorithmes d'analyse efficaces, on peut limiter le nombre de trous maximal à 1 avec une perte négligeable de pouvoir expressif.

3.2.2. *Nombre maximal de composantes connexes de trous*

Ce paramètre a été introduit par Nivre (2006), toujours dans un but d'augmenter la performance des algorithmes d'analyse sans réduire trop le pouvoir d'expression.

**Définition 2.** *Une composante connexe de trous d'une dépendance* `H` $\rightarrow$ `D` *est un ensemble maximal de nœuds situés entre* `H` *et* `D` *qui ne sont pas dans la projection de* `H` *et qui sont connectés les uns aux autres.*

Pour l'annotation SUD de la figure 1, la dépendance (`donné -[comp=obj]`$\rightarrow$ `que`) comporte deux trous {*je*} et {*ai*} mais une seule composante connexe de trous car *je* et *ai* sont liés par une dépendance.

Nous avons conçu un programme Python qui, pour chaque dépendance d'un corpus, détermine la liste de ses composantes connexes. Ces composantes sont identifiées

| corpus | SUD-GSD | UD-GSD | UD-SEQUOIA | total |
|---|---|---|---|---|
| nb. de dépendances | 356 393 | 356 393 | 62 706 | 775 492 |
| nb. de dépendances non projectives | 1 547 (0,47 %) | 679 (0,19 %) | 73 (0,12 %) | 1 924 (0,25 %) |
| dépendances à 0 composante connexe de trous | 354 846 | 355 714 | 62 633 | 773 193 |
| dépendances à 1 composante connexe de trous | 1 531 | 667 | 71 | 2 269 |
| dépendances à 2 composantes connexes de trous | 15 | 11 | 2 | 28 |
| dépendances à 3 composantes connexes de trous | 1 | 1 | 0 | 2 |
| nb. d'arbres syntaxiques | 16 341 | 16 341 | 3 099 | 35 781 |
| nb. d'arbres non projectifs | 1 392 (8,52 %) | 653 (4,00 %) | 66 (2,13 %) | 2 111 (5,90 %) |
| arbres à 0 composante connexe de trous max. | 14 949 | 15 688 | 3 033 | 33 670 |
| arbres à 1 composante connexe de trous max. | 1 376 | 641 | 64 | 2 081 |
| arbres à 2 composantes connexes de trous max. | 15 | 11 | 2 | 28 |
| arbres à 3 composantes connexes de trous max. | 1 | 1 | 0 | 2 |

**Tableau 2.** *Statistiques sur le nombre de composantes connexes de trous par dépendances*

par leurs racines. Ces racines sont faciles à déterminer car elles se caractérisent comme un nœud d'un trou dont le gouverneur est extérieur à la dépendance considérée.

Le tableau 2 récapitule les résultats trouvés sur nos trois corpus. Il montre qu'on peut limiter le nombre maximal de composantes connexes à 1. Nivre (2006) a établi expérimentalement sur les treebanks DANISH DEPENDENCY TREEBANK et PRAGUE DEPENDENCY TREEBANK qu'avec cette limite, on peut obtenir des algorithmes d'analyse linéaires en temps, qui excluent moins de 2 % des solutions. Il reste à faire le même type de mesure sur nos corpus du français.

### 3.3. *La non-projectivité, un phénomène local*

Considérons maintenant la caractérisation de la non-projectivité comme croisement de deux dépendances. Cette configuration n'est pas à proprement parler locale au sens où elle constituerait un graphe fini connexe, quand on ne considère que les dépendances syntaxiques : elle ne fait apparaître aucun lien entre les deux dépendances qui se croisent. Or, il faut se rappeler que ces deux dépendances font partie d'un même arbre donc les chemins qui mènent de $\mathtt{D}_1$ et de $\mathtt{D}_2$ à la racine de l'arbre se rencontrent

| **Chaîne reliant les deux dépendances** | **DIST( H, $D_1$)** | **DIST( H, $D_2$)** | **dépendance non projective** |
|---|---|---|---|
| $H_1$ $^{+}\leftarrow$ H $\rightarrow^{+}H_2$ | > 1 | > 1 | $H_1 \rightarrow D_1$ , $H_2 \rightarrow D_2$ |
| $H_1 \rightarrow^{+} H_2$ | 1 | > 1 | $H_2 \rightarrow D_2$ |
| $D_1 \rightarrow^{+} H_2$ | 0 | > 1 | $H_2 \rightarrow D_2$ |
| $H_2 \rightarrow^{+} H_1$ | > 1 | 1 | $H_1 \rightarrow D_1$ |
| $D_2 \rightarrow^{+} H_1$ | > 1 | 0 | $H_1 \rightarrow D_1$ |

**Tableau 3.** *Les cinq configurations formées par deux dépendances* $H_1 \rightarrow D_1$ *et* $H_2 \rightarrow D_2$ *qui se croisent*

nécessairement en un nœud que nous noterons H. Pour déterminer dans quelle mesure la non-projectivité est un phénomène local, il est intéressant d'étudier les distances de $D_1$ et de $D_2$ à H et de voir à quel point elles sont bornées. Dans un arbre, la distance d'un nœud N à un de ces ancêtres A, notée DIST(A, N), est le nombre de dépendances formant le chemin de A à N.

Pour mesurer le caractère local d'une dépendance non projective, Havelka (2007) définit une distance qui a un rapport direct avec les distances DIST(H, $D_1$) et DIST(H, $D_2$). Sa distance n'est pas attachée à une paire de dépendances qui se croisent mais à une dépendance non projective. Considérons-en une quelconque $H_1 \rightarrow D_1$. Havelka définit le *type de niveau* de cette dépendance comme étant le maximum de DIST(H, $D_1$) − DIST(H, $D_2$), pour $H_2 \rightarrow D_2$ étant une dépendance qui croise la première, et H étant le premier ancêtre commun à $D_1$ et $D_2$. L'objectif de Havelka est de concevoir des algorithmes efficaces de détermination de dépendances non projectives alors que le nôtre est d'exhiber des configurations locales de dépendances qui se croisent.

Selon les positions de $D_1$ et de $D_2$ par rapport à H, il y a cinq configurations possibles. Pour chacune d'elles, nous indiquons la forme de la chaîne reliant les deux dépendances qui se croisent et les valeurs possibles de DIST(H, $D_1$) et de DIST(H, $D_2$). On peut même déterminer la ou les dépendances responsables de la non-projectivité. Le tableau 3 décrit ces cinq configurations, en indiquant laquelle des deux dépendances concernées est nécessairement non projective[12]. Théoriquement, la distance DIST(H, $D_1$) ou DIST(H, $D_2$) est non bornée. Nous nous proposons de voir ce qu'il en est sur corpus. Pour cela, nous avons conçu un programme qui prend en entrée un corpus annoté en dépendances syntaxiques dans un format *conll* et qui retourne un fichier contenant toutes les dépendances qui se croisent en indiquant pour chaque croisement la valeur de DIST(H, $D_1$) et celle de DIST(H, $D_2$). En plus, sont affichées certaines statistiques sur le corpus relatives à la non-projectivité.

On a appliqué le programme aux trois corpus du français SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA. Le tableau 4 en récapitule les résultats. Le rôle déterminant du format est confirmé par ces statistiques : il y a 2,6 fois plus de croisements dans SUD_FRENCH-GSD que dans UD_FRENCH-GSD.

12. Nous rappelons que la relation $\rightarrow^{+}$ est la clôture transitive de la relation de dépendance.

| **corpus** | **SUD-GSD** | **UD-GSD** | **UD-SEQUOIA** | **total** |
|---|---|---|---|---|
| nb. de tokens | 416 740 | 416 740 | 73 666 | 907 146 |
| nb. de croisements | 3 283 (0,79 %) | 1 242 (0,30 %) | 107 (0,15 %) | 4 632 (0,51 %) |
| croisements avec $\mathtt{H}_1 \ ^+\!\leftarrow \mathtt{H} \rightarrow^+\!\mathtt{H}_2$ | 5 | 3 | 0 | 8 |
| croisements avec $\mathtt{H}_1 \rightarrow^+\!\mathtt{H}_2$ | 1 205 | 609 | 42 | 1 856 |
| croisements avec $\mathtt{D}_1 \rightarrow^+\!\mathtt{H}_2$ | 282 | 283 | 5 | 570 |
| croisements avec $\mathtt{H}_2 \rightarrow^+\!\mathtt{H}_1$ | 716 | 114 | 13 | 743 |
| croisements avec $\mathtt{D}_2 \rightarrow^+\!\mathtt{H}_1$ | 1 075 | 233 | 47 | 1 355 |
| maximum de DIST(H, $\mathtt{D}_1$) | 5 | 3 | 3 | 5 |
| maximum de DIST(H, $\mathtt{D}_2$) | 5 | 4 | 3 | 5 |

**Tableau 4.** *Récapitulation des croisements pour trois corpus du français*

| distance à H | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| distribution de DIST(H, $\mathtt{D}_1$) | 570 | 1 856 | 2 028 | 164 | 13 | 1 |
| distribution de DIST(H, $\mathtt{D}_2$) | 1 355 | 843 | 2 264 | 150 | 13 | 5 |

**Tableau 5.** *Répartition des dépendances qui se croisent par distance à l'ancêtre commun* H

Une seconde constatation est que la configuration $\mathtt{H}_1 \ ^+\!\leftarrow \mathtt{H} \rightarrow^+\!\mathtt{H}_2$ est très rare. Parmi les 5 configurations possibles, c'est la seule qui donne lieu à une imbrication de sous-arbres disjoints. On peut donc avec une perte négligeable d'expressivité utiliser des algorithmes d'analyse qui prennent en compte la non-imbrication de sous-arbres disjoints pour plus d'efficacité (Kuhlmann et Nivre, 2006 ; Corro *et al.*, 2016).

Venons-en maintenant à la question principale qui nous préoccupe : dans quelle mesure la configuration de deux dépendances qui se croisent est-elle locale ? Pour y répondre, il faut examiner les limites supérieures de DIST(H, $\mathtt{D}_1$) et de DIST(H, $\mathtt{D}_2$). Sur les trois corpus étudiés, ces limites sont toutes les deux 5, mais en plus, elles sont rarement atteintes. La tableau 5 est révélateur à ce sujet.

Il existe seulement six phrases pour lesquelles la limite 5 de DIST(H, $\mathtt{D}_1$) ou de DIST(H, $\mathtt{D}_2$) est atteinte. Elles sont dans le corpus SUD_FRENCH-GSD.

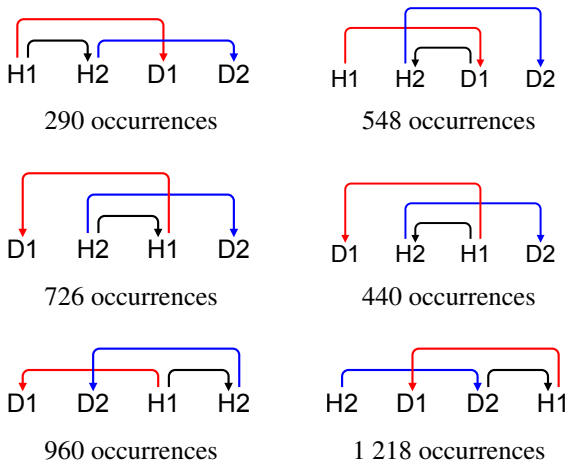


**Figure 3.** *Les six motifs de croisement de dépendances les plus fréquents*

Détaillons ce qu'il en est sur l'une d'elles, la phrase *fr-ud-train_02164*.

> *Ce sont là, avec d'autres, des actes clairement hostiles à lesquels (auxquels) il n'a pas été jugé utile de répondre pour l'instant.*

Les deux dépendances qui se croisent sont *actes* → *a* et *à* ← *répondre*. Elles sont liées par le chemin *a* → *été* → *jugé* → *de* → *répondre* → *à*. Ce chemin est bien de longueur 5. On pourrait penser que sa longueur exceptionnelle rend la compréhension de la phrase difficile mais on peut remarquer que ce chemin suit l'ordre de lecture de la phrase, ce qui explique peut-être que la phrase est aisément compréhensible.

En nous limitant à des motifs où la chaîne qui va de $\mathtt{D}_1$ à $\mathtt{D}_2$ a une longueur au maximum de 4, on couvre 4 611 des 4 632 croisements de dépendances. Ces motifs sont au nombre de 11 et si on les combine avec les quatre façons d'ordonner les deux dépendances qui se croisent, on obtient 44 motifs. Sur ces 44 motifs, il y en a seulement 21 dont on rencontre des occurrences dans au moins un des trois corpus et parmi ces 21, il n' y en a que 11 qui ont au moins 49 occurrences, tous les autres ont moins de 15 occurrences. Enfin sur ces 11, 6 se dégagent nettement, couvrant 4 144 croisements sur 4 632. La figure 3 présente ces six motifs[13].

Même si l'étude que nous avons menée est dépendante des trois corpus sur lesquels elle a porté, la conclusion est que la non-projectivité en français est un phénomène très local. Nous sélectionnons les 11 motifs les plus fréquents qui vont nous permettre de pousser plus à fond l'étude en considérant maintenant la dimension linguistique.

13. Ces six motifs seront illustrés par des exemples dans la partie 4 et dans chaque exemple, pour le croisement concerné, la dépendance nœuds $\mathtt{H}_1 \rightarrow \mathtt{D}_1$ sera en rouge et la dépendance $\mathtt{H}_2 \rightarrow \mathtt{D}_2$ en bleu.

## 4. Étude linguistique des dépendances non projectives en français

Il s'agit maintenant de réintégrer les étiquettes des fonctions syntaxiques dans les dépendances formant les 11 motifs qui ont été exhibés dans la section précédente, afin de dégager les sources linguistiques de la non-projectivité en français. Nous commencerons par présenter l'outil informatique GREW-MATCH que nous avons utilisé pour notre étude puis nous donnerons les résultats de l'étude elle-même.

### 4.1. *L'outil de recherche automatique de motifs dans un graphe* GREW-MATCH

GREW-MATCH[14] est un outil qui permet de retrouver dans un graphe toutes les occurrences d'un motif donné.

```
pattern{ H1 -> D1; H2 -> D2; D2 -> H1;
         H2 << D1; D1 << D2; D2 << H1 }
without{ D1[upos=PUNCT] }
without{ D2-[1= comp]-> H1; D2[upos=AUX|VERB]; D1[upos=PRON] }
without{ H2 -[mod@relcl]-> D2; D1[PronType=Rel] }
without{ H2 -[mod@relcl]-> D2; D1 -> P; P[PronType=Rel] }
without{ H2 -[mod@relcl]-> D2; D1 -> D; D -> P; P[PronType=Rel] }
```

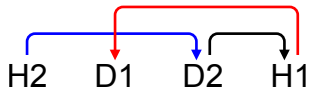


**Figure 4.** *Exemple de motif* GREW

La figure 4 présente un motif exprimant une configuration de deux dépendances $H_1 \rightarrow D_1$ et $H_2 \rightarrow D_2$ qui se croisent. Au-dessus, vous avez la définition du motif dans la syntaxe de GREW et au-dessous la traduction graphique de sa partie *pattern*. Le mot-clé *pattern* permet de décrire le motif sous forme d'une suite de déclarations et contraintes élémentaires séparées par un point-virgule. La première ligne contient les déclarations de trois dépendances et la suivante exprime des contraintes d'ordre entre les nœuds déclarés avec les dépendances. Chaque mot-clé *without* introduit une contrainte négative. Par exemple, le premier *without* exclut que $D_1$ soit un signe de ponctuation [15]. On peut mettre plusieurs *without* dans un motif.

On utilise GREW-MATCH de façon itérative : pour un motif de départ, on observe les occurrences retournées et on identifie un sous-motif linguistiquement pertinent qui revient régulièrement. On itère ensuite les recherches avec le motif de départ et en

14. `http://match.grew.fr`

15. Nous avons écarté de notre étude les dépendances impliquant des signes de ponctuation et les autres nœuds déclarés dans *pattern* ne peuvent pas l'être car un signe de ponctuation n'est jamais gouverneur d'une dépendance.

excluant (*without*) tous les sous-motifs identifiés aux étapes précédentes. Le motif de la figure 4 illustre le type de requête que l'on obtient après quelques itérations.

Dans la première étape, on applique le motif formé seulement du champ *pattern* et du premier *without* au corpus SUD_FRENCH-GSD ; on trouve 955 occurrences du motif dans le corpus, correspondant à 955 croisements. Un sous-motif revenant souvent caractérise la montée de clitiques. Pour l'exclure, on ajoute le second *without* de la figure 4, qui signifie que nous ne voulons pas qu'un complément $\texttt{H}_1$ d'un verbe $\texttt{D}_2$ ait lui-même un dépendant $\texttt{D}_1$ qui soit un pronom personnel, ce pronom se situant avant le verbe $\texttt{D}_2$. En appliquant le motif enrichi de cette contrainte, on ne trouve plus que 284 occurrences dans le corpus. Donc 671 étaient dues à la montée de clitiques.

En observant un échantillon de ces 284 occurrences, on s'aperçoit que les croisements sont souvent dus à l'extraction profonde de propositions relatives. Nous employons le terme *extraction profonde* pour indiquer que le syntagme extrait n'est pas directement dépendant de la tête de la proposition relative. Pour exclure ce phénomène et en rechercher d'autres nous ajoutons les trois derniers *without* de la figure 4. Ils expriment que $\texttt{H}_2$ est l'antécédent d'un pronom relatif, que $\texttt{D}_2$ est la tête de la relative et que le pronom relatif repéré par le trait `PronType=Rel` dépend plus ou moins directement de $\texttt{H}_1$, qui dépend, lui, directement de $\texttt{D}_2$. La dépendance plus ou moins directe du pronom relatif explique la nécessité de trois *without*, qui, ensemble, expriment le rejet de l'extraction profonde d'une relative. Lorsque l'on applique le motif complet de la figure 4, on ne trouve plus que 44 occurrences de ce motif dans le corpus. Cela signifie que 240 croisements provenaient d'une extraction profonde d'une relative.

On poursuit ce processus jusqu'à ce qu'il ne reste plus qu'une dizaine d'occurrences qui peuvent correspondre à des phénomènes extrêmement rares, mais le plus souvent mettent en exergue des erreurs d'annotation.

C'est en appliquant cette méthode qu'a été menée l'étude linguistique des dépendances non projectives dont nous allons maintenant présenter les résultats.

### 4.2. *Les sources de la non-projectivité en français*

Dans la section précédente, nous avions montré l'existence de 11 motifs principaux de manifestation de la non-projectivité dans nos trois corpus étudiés. À l'aide de GREW-MATCH, nous avons appliqué la méthode qui vient d'être présentée à chacun des trois corpus en partant de chacun des 11 motifs. Ainsi, nous avons pu exhiber quatre phénomènes principaux sources de non-projectivité, qui couvrent 97 % des croisements. Le tableau 6 récapitule ces résultats en indiquant pour chaque phénomène le nombre de croisements auquel il donne lieu par corpus[16]. Nous allons maintenant détailler ces quatre phénomènes les uns après les autres.

16. Un motif correspond généralement à plusieurs phénomènes et un phénomène correspond à plusieurs motifs. Le lecteur trouvera un tableau récapitulant cette correspondance en ligne (`https://nakala.fr/10.34847/nkl.ce3cmi1q`).

| **phénomène linguistique** | **SUD-GSD** | **UD-GSD** | **UD-SEQUOIA** | **total** |
|---|---|---|---|---|
| montée de clitiques | 2 154 (67 %) | 192 (16 % ) | 23 | 2 369 (53 % ) |
| extraction profonde | 496 (15 %) | 196 (17 %) | 48 | 740 (16 %) |
| minimisation de la longueur des dépendances | 290 (9 %) | 255 (22 %) | 30 | 575 (13 %) |
| couples de mots dépendants distants | 274 (9 %) | 532 (45 %) | 0 | 806 (18 %) |
| total | 3 214 | 1 175 | 101 | 4 490 |

**Tableau 6.** *Sources linguistiques de la non-projectivité dans les corpus du français étudiés*

### 4.2.1. *La montée de clitiques*

Les pronoms clitiques sont accolés aux verbes dont ils sont les compléments mais quand ces verbes sont précédés d'auxiliaires, les clitiques montent devant les auxiliaires (Abeillé et Godard, 2001). La figure 5 illustre cette montée.

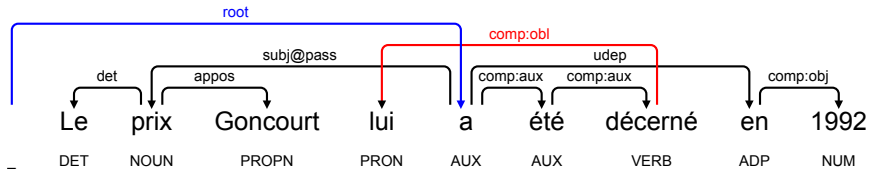


**Figure 5.** SUD_FRENCH-GSD *fr-ud-train_06154 : Le prix Goncourt lui a été décerné en 1992*

Dans cette phrase, la montée du clitique *lui* devant l'auxiliaire *a* entraîne nécessairement un croisement dans l'annotation SUD entre la dépendance ( `décerné -[comp:obl]→ lui)` et la dépendance (`_ -[root]→ a)` dans l'arbre complété. Dans le corpus SUD_FRENCH-GSD, on trouve 612 occurrences de tels croisements.

Ces 612 occurrences représentent 28 % seulement des croisements qui, dans le tableau 6, sont considérés comme résultant d'une montée de clitique, car cette montée entraîne aussi des croisements secondaires. Sur la figure 5, la montée du clitique *lui* provoque un croisement avec les dépendants à gauche de l'auxiliaire. Dans l'exemple, il s'agit du croisement avec la dépendance (`a -[subj@pass]→ prix`). Dans le corpus SUD_FRENCH-GSD il y a 819 croisements de ce type, soit 38 % des croisements provoqués par une montée de clitique. Ils sont plus nombreux que les croisements principaux car il peut y avoir plusieurs dépendants à gauche d'un même auxiliaire. Ces dépendants à gauche sont essentiellement des sujets (569), des modificateurs de phrases (169) et des conjonctions de coordination (79).

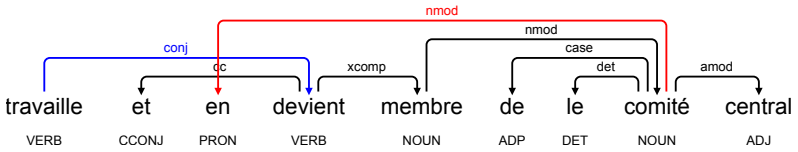


**Figure 6.** UD_FRENCH-GSD *fr-ud-train_07141 : [Smrkovský] travaille [pour la résistance communiste allemande] et en devient [finalement] membre du comité central*

La montée d'un clitique peut provoquer aussi des croisements secondaires avec les dépendants à droite de l'auxiliaire. Dans l'exemple, il s'agit du croisement de (`décerné -[udep]→ lui`) avec (`a -[udep]→ en`). Dans le corpus SUD_FRENCH-GSD, il y a 549 croisements de ce type, soit 25 % des croisements provoqués par une montée de clitique. Les dépendants à droite correspondant à ces croisements sont principalement des modificateurs de phrases (426) et des têtes de propositions coordonnées (78).

Dans le format UD, les auxiliaires ne sont pas la tête du complexe qu'ils forment avec le verbe principal donc la montée des clitiques devant les auxiliaires ne provoque aucun croisement. Néanmoins, certains clitiques ne dépendent pas directement du verbe auquel ils sont accolés mais d'un argument de ce verbe. Ce phénomène peut provoquer des croisements. Cela concerne essentiellement le clitique *en* lorsqu'il dépend de l'objet du verbe ou de l'attribut du sujet ou de l'objet. La figure 6[17] montre que le chemin de dépendances du verbe vers le clitique *en* qui lui est accolé peut être plus ou moins long, ici : *devient* → *membre* → *comité* → *en*. Ce phénomène pour le clitique *en* est responsable de respectivement 161, 185 et 21 croisements dans SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, soit respectivement 7 %, 96 % et 91 % des croisements résultant d'une montée de clitique.

### 4.2.2. *L'extraction profonde*

Les propositions relatives ou interrogatives donnent lieu dans beaucoup de cas à l'extraction d'un syntagme, mais quand ce syntagme est directement dépendant de la tête de la proposition où a lieu l'extraction, celle-ci n'entraîne pas de non-projectivité dans les dépendances. Pour que cette non-projectivité se produise, il faut que le syntagme extrait dépende d'un élément qui n'est pas la tête de la proposition relative ou interrogative. C'est alors que nous parlons d'*extraction profonde*.

Toute extraction profonde n'entraîne pas de non-projectivité. Il faut pour cela que le gouverneur du syntagme extrait se situe après la tête de la proposition rela-

17. Pour simplifier la présentation de l'annotation, nous ignorons certains passages de la phrase non essentiels pour notre propos ; ceux-ci sont marqués entre crochets dans l'énoncé de la phrase.

tive. La figure 7 montre un exemple d'extraction profonde qui entraîne de la non-projectivité. C'est *dont* qui est extrait de la relative comme complément du nom

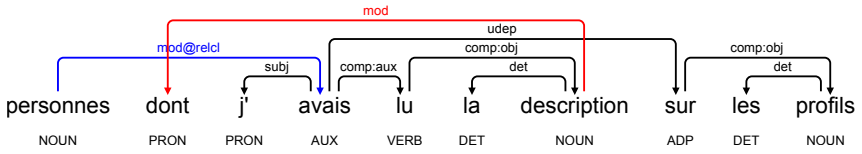


**Figure 7.** SUD_FRENCH-GSD *fr-ud-train_11296 : [j'ai envoyé des messages à plusieurs] personnes dont j'avais lu la description sur les profils [donnés par unicis]*

*description*. Ce gouverneur est bien situé après à la tête *avais* de la relative. La dépendance correspondant à l'extraction (`description -[mod]→ dont`) croise donc nécessairement celle de l'antécédent vers la tête de la relative, (`personnes -[mod@relcl]→ avais`). Dans les corpus SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, nous trouvons respectivement 333, 153 et 35 occurrences de ce type de croisement, soit 11 %, 5 % et 6 % du nombre total de relatives et interrogatives.

Souvent, l'extraction profonde entraîne des croisements secondaires avec les dépendants à droite de la tête de la relative ou de l'interrogative. C'est ce que nous constatons sur l'exemple de la figure 7. La dépendance correspondant à l'extraction (`description -[mod]→ dont`) croise la dépendance (`avais -[mod]→ sur`). Dans les corpus SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, nous rencontrons respectivement 149, 34 et 10 occurrences de tels croisements, soit 30 %, 17 % et 21 % du nombre total de croisements résultant d'extractions profondes.

Ce phénomène d'extraction profonde pour le français a été étudié par Candito et Seddah (2012a) sous le nom de *dépendances à longue distance effectives*. Ils y incluent aussi les clitiques qui dépendent d'arguments du verbe auquel ils sont accolés et dont nous avons parlé précédemment. Ils ont décrit relativement précisément ce phénomène sur deux treebanks, FRENCH TREEBANK et SEQUOIA, en le quantifiant. Ils se fondent sur une annotation LFG des corpus en utilisant les chemins fonctionnels propres à ce formalisme pour retrouver les dépendances à longue distance. Même si notre méthode est différente, nous retrouvons sur le corpus SEQUOIA les mêmes résultats en termes de croisements.

### 4.2.3. *La minimisation de la longueur des dépendances*

Ferrer Cancho (2006) a mis en évidence le fait que la meilleure façon d'ordonner les nœuds d'un arbre de dépendances pour minimiser la longueur des dépendances est de le faire de façon projective. Et la minimisation de la longueur des dépendances, MLD par la suite, aide à la compréhension comme l'a étudié Liu (2008). Cette minimisation peut être interprétée aussi comme une minimisation du flux des dépendances

dans un arbre de dépendances totalement ordonné, ce qui rend mieux compte de l'aspect cognitif de la question, comme l'ont montré Kahane et Yan (2019).

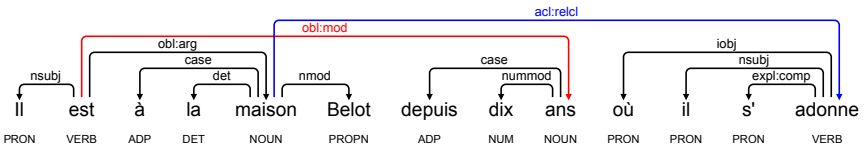


**Figure 8.** UD_FRENCH-SEQUOIA *annodis.er_00519 : Il est à la maison Belot depuis dix ans où il s'adonne [à la belote avec son copain Pierre Brungard]*

Si on met de côté les phénomènes de montée des clitiques et d'extraction profonde, qui viennent d'être présentés, on peut alors se demander pourquoi d'autres dépendances non projectives sont présentes dans les corpus. C'est que l'ordre linéaire des mots est soumis à d'autres contraintes qui peuvent aller à l'encontre du principe de la MLD. En particulier, la structure communicative avec la division entre thème et rhème et la focalisation impose ses propres contraintes sur l'ordre des mots. Il s'agit alors d'appliquer la MLD à l'intérieur de ces contraintes, ce qui entraîne parfois des croisements.

C'est en particulier le cas pour la phrase présentée sur la figure 8. Le verbe *est* a deux compléments *à la maison Belot* et *depuis dix ans* mais le premier complément est modifié par la relative *où il s'adonne à la belote avec son copain Pierre Brungard*. Or, cette relative n'est pas accolée à son antécédent comme c'est le cas habituellement. Elle en est séparée par *depuis dix ans*. Cet enchevêtrement entraîne un croisement de dépendances. On peut se demander pourquoi la phrase n'est pas plutôt structurée selon l'une des deux alternatives projectives suivantes : *Il est à la maison Belot où il s'adonne à la belote avec son copain Pierre Brungard depuis dix ans* et *Il est depuis dix ans à la maison Belot où il s'adonne à la belote avec son copain Pierre Brungard*. La première alternative, même si elle raccourcit la dépendance (`maison -[acl:relcl]→ adonne`), allonge considérablement la dépendance (`est -[obl:mod]→ ans`), si bien que cela entraîne une ambiguïté : on peut comprendre que c'est depuis dix ans qu'il s'adonne à la belote. La seconde n'est pas satisfaisante du point de vue de la structure communicative car l'information nouvelle qui veut être mise en avant c'est que cela fait dix ans qu'il est à la maison Belot et pas que c'est à la maison Belot qu'il est. La seule façon de concilier la structure communicative voulue avec la minimisation de la longueur des dépendances est la phrase de la figure 8. On retrouve ce type d'enchevêtrement de compléments ou modificateurs dans respectivement 182, 143 et 23 occurrences des treebanks SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, soit 63 %, 56 % et 77 % du nombre total de croisements provenant de la minimisation de la longueur des dépendances.

Il est une autre configuration qui se rapporte au même phénomène et qui est illustrée par la figure 9 : le sujet d'un verbe peut voir un de ses modificateurs ou une apposition rejetés après le verbe. Dans les corpus SUD_FRENCH-GSD, UD_FRENCH-

GSD, UD_FRENCH-SEQUOIA, on trouve respectivement 71, 69 et 4 occurrences de cette configuration, soit 24 %, 27 % et 13 % du nombre total de croisements provenant de la minimisation de la longueur des dépendances.

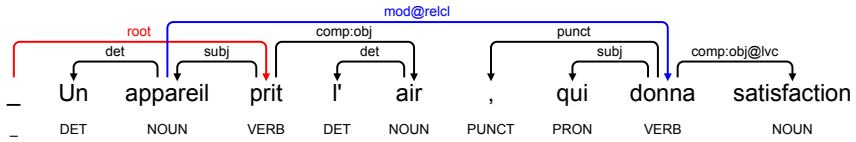


**Figure 9.** SUD_FRENCH-GSD *fr-ud-train_13250 : Un [second] appareil prit l'air [le 20 mars 1999], qui donna satisfaction*

### 4.2.4. *Les couples de mots dépendants distants*

Les constructions comparatives en français utilisent en général des couples de mots dépendants dont le premier est un adverbe (*plus*, *moins*, *aussi*, *autant*, *davantage*...) ou un adjectif (*autre*, *même*...) et le second la conjonction de subordination *que*. Il est en de même avec les constructions consécutives (*si*... *que*, *tel*... *que*, *tellement*... *que*...). Dans les deux types de constructions, les deux mots dépendants peuvent être distants l'un de l'autre et c'est cela qui est source de non-projectivité. On retrouve des couples de mots dépendants potentiellement distants pour d'autres constructions : *premier*... *à*, *à peine* ... *que* par exemple.

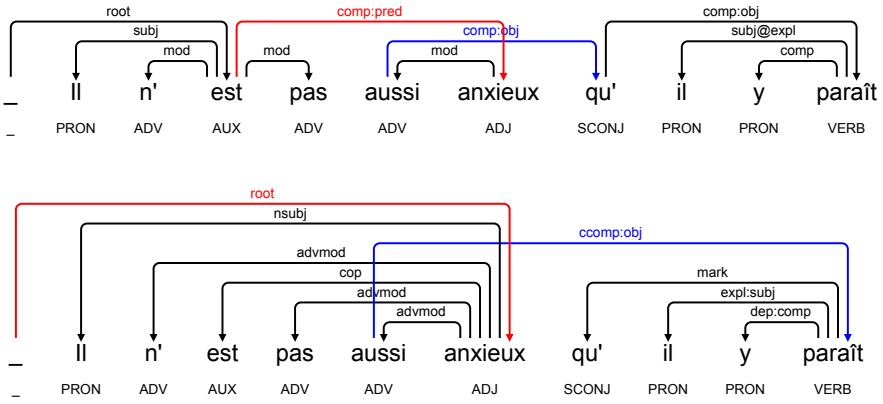


**Figure 10.** SUD_FRENCH-GSD *et* UD_FRENCH-GSD *fr-ud-dev_00420 : Il n'est pas aussi anxieux qu'il y paraît*

La figure 10 montre un exemple avec la construction *aussi* ... *que*. Le schéma supérieur présente l'annotation SUD de la phrase qui fait apparaître le croise-

ment de la dépendance (`aussi -[comp:obj]→ qu'`) avec (`est -[comp:pred]→ anxieux`). Dans l'annotation UDprésentée dans la partie inférieure de la figure 10, la relation entre les mots distants *aussi* et *qu'* est exprimée de façon indirecte par la dépendance (`aussi -[ccomp:obj]→ paraît`), qui se croise cette fois avec (`_ -[root]→ anxieux`).

Dans les corpus SUD_FRENCH-GSD, UD_FRENCH-GSD, on trouve respectivement 197 et 194 occurrences de ces croisements. Dans UD_FRENCH-SEQUOIA, nous n'en trouvons aucun car il a été fait un choix différent pour la source de la dépendance exprimant la relation entre les mots distants. La source de cette dépendance n'est pas le premier mot du couple considéré mais son gouverneur. Le lecteur pourra aisément vérifier que dans l'annotation UD de la figure 10, si on déplace la source de la dépendance (`aussi -[ccomp:obj]→ paraît`) sur *anxieux*, on supprime la non-projectivité.

Comme le montre la figure 10, la dépendance principale $H_2 \rightarrow D_2$ (en bleu) induit des croisements secondaires avec des dépendances issues de $D_1$ (la cible de la dépendance en rouge). Dans notre exemple, on n'en trouve aucun pour SUD mais quatre pour UD. Cette constatation est plus générale puisque sur l'ensemble des corpus SUD_FRENCH-GSD et UD_FRENCH-GSD, il y a respectivement 77 et 338 croisements secondaires induits par la dépendance impliquant les mots distants.

## 5. Conclusion

Notre étude sur les dépendances non projectives dans des corpus du français a fait apparaître la complexité limitée de la non-projectivité, que ce soit en nombre de trous ou de composantes connexes de trous. Elle a fait apparaître aussi le caractère très local du croisement de dépendances. Ces deux résultats sont indépendants du schéma d'annotation. En revanche, l'étude a montré que la fréquence de croisement était en partie déterminée par le schéma d'annotation.

Ensuite, elle a mis en évidence quatre sources linguistiques de non-projectivité : la montée de clitiques, l'extraction profonde, la minimisation de la longueur des dépendances et les couples de mots dépendants distants. Ces phénomènes ont déjà été largement étudiés par le passé (Abeillé et Godard, 2001 ; Candito et Seddah, 2012a ; Ferrer Cancho, 2006 ; Liu, 2008), mais en général indépendamment les uns des autres et pas sous l'angle de leur responsabilité dans la non-projectivité. L'étude que nous avons menée visait à l'exhaustivité en mettant en évidence toutes les sources linguistiques de non-projectivité dans les corpus considérés. Les quatre sources exhibées couvrent 97 % des croisements de dépendances dans les corpus considérés. L'intérêt de l'étude est aussi que les différents phénomènes ont été quantifiés. La seule autre étude à notre connaissance visant à déterminer l'ensemble des sources de non-projectivité dans un corpus du français est celle de Botalla (2014) menée sur le treebank RHAPSODIE, mais elle reste qualitative et elle considère différentes manifestations de la minimisation de la longueur des dépendances sans l'envisager dans toute sa généralité.

Le fait d'avoir considéré deux schémas d'annotation syntaxique très différents met en évidence que la non-projectivité dépend parfois des choix qui sont faits pour la tête des constituants. Plus précisément pour SUD et UD la divergence porte sur la tête des groupes prépositionnels, des propositions subordonnées et des noyaux verbaux avec auxiliaires. La non-projectivité liée à l'extraction profonde et à la minimisation de la longueur des dépendances est pour une bonne part indépendante de ces choix de têtes, alors que la non-projectivité liée à la montée des clitiques est due la plupart du temps au choix de l'auxiliaire comme tête dans le couple qu'il forme avec le verbe. Pour ce qui est des couples de mots dépendants distants, ce qui joue n'est pas le choix de la tête d'un constituant mais celui de la source de la dépendance qui établit une relation plus ou moins directe entre les deux mots distants, comme cela est expliqué avec l'exemple de la figure 10.

Enfin, l'intérêt de l'étude effectuée est qu'elle utilise une méthode universelle, valable quelle que soit la langue, quel que soit le schéma d'annotation et quel que soit le type de corpus. C'est vrai pour la première phase d'étude topologique. L'aboutissement de cette première phase est de mettre en évidence un nombre limité de motifs finis couvrant l'essentiel des croisements de dépendance. Bien entendu, ces motifs dépendent de la langue concernée, mais ils sont le point de départ d'une étude exhaustive et quantifiée à l'aide de l'outil GREW-MATCH des phénomènes linguistiques qui sont source de non-projectivité. Comme prolongement immédiat de notre étude, il serait intéressant d'appliquer la méthode au treebank de l'oral SPOKEN[18]. Cela permettrait, par comparaison avec l'étude présentée ici sur des treebanks de l'écrit, de mettre en évidence les points communs et les spécificités de l'oral.

Remerciements

Merci à Sylvain Kahane, Bruno Guillaume et aux relecteurs anonymes pour leurs commentaires pertinents.

## 6. Bibliographie

Abeillé A., Clément L., Liégeois L., « Un corpus annoté pour le français : le French Treebank », *TAL*, vol. 60, p. 19-43, 2019.

Abeillé A., Godard D., « Deux types de prédicats complexes dans les langues romanes », *Linx. Revue des linguistes de l'université Paris X Nanterre*, nº 45, p. 167-175, 2001.

Béchet D., Lacroix O., « CDGFr, un corpus en dépendances non-projectives pour le français », *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, p. 206-212, 2015.

Bloomfield L., *Language*, New-york, 1933.

18. `https://github.com/surfacesyntacticud/SUD_French-Spoken`

Bonfante G., Guillaume B., Perrier G., *Application de la réécriture de graphes au traitement automatique des langues*, vol. 1 of *Série Logique, linguistique et informatique*, ISTE editions, 2018.

Botalla M.-A., *« Analyse du flux de dépendance dans un corpus de français oral annoté en micro-syntaxe »*, Master's thesis, Université Paris III Sorbonne Nouvelle, 2014.

Candito M.-H., Crabbé B., Denis P., Guérin F., « Analyse syntaxique statistique du français : des constituants aux dépendances », *TALN 2009*, Senlis, France, 2009.

Candito M., Seddah D., « Effectively long-distance dependencies in French : annotation and parsing evaluation », *TLT 11 - The 11th International Workshop on Treebanks and Linguistic Theories*, Lisbon, Portugal, 2012a.

Candito M., Seddah D., « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France, 2012b.

Corro C., Le Roux J., Lacroix M., Rozenknop A., Calvo R. W., « Dependency parsing with bounded block degree and well-nestedness via Lagrangian relaxation and branch-and-bound », *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 355-366, 2016.

Ferrer Cancho R., « Why do syntactic links not cross ? », *Europhysics Letters*, vol. 76, n° 6, p. 1228-1235, 2006.

Fitialov S. J., « O modelirovanii sintaksisa v strukturnoj lingvistike », *Problemy strukturnoj lingvistiki, Moskva*p. 100-114, 1962.

Gaifman H., « Dependency systems and phrase-structure systems », *Information and control*, vol. 8, n° 3, p. 304-337, 1965.

Gerdes K., Guillaume B., Kahane S., Perrier G., « SUD or Surface-Syntactic Universal Dependencies : An annotation scheme near-isomorphic to UD », *Universal Dependencies Workshop 2018*, Brussels, Belgium, 2018.

Gerdes K., Guillaume B., Kahane S., Perrier G., « Improving Surface-syntactic Universal Dependencies (SUD) : surface-syntactic relations and deep syntactic features », *TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories*, Paris, France, 2019.

Gómez-Rodríguez C., Sartorio F., Satta G., « A polynomial-time dynamic oracle for non-projective dependency parsing », *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 917-927, 2014.

Guillaume B., de Marneffe M.-C., Perrier G., « Conversion et améliorations de corpus du français annotés en Universal Dependencies », *Traitement Automatique des Langues*, vol. 60, n° 2, p. 71-95, 2019.

Hajicová E., Havelka J., Sgall P., Veselá K., Zeman D., « Issues of Projectivity in the Prague Dependency Treebank. », *Prague Bull. Math. Linguistics*, vol. 81, p. 5-22, 2004.

Harper K. E., Hays D. G., *The use of machines in the construction of a grammar and computer program for structural analysis*, Rand Corporation, 1959.

Havelka J., Mathematical Properties of Dependency Trees and their Application to Natural Language Syntax, PhD thesis, Univerzita Karlova, Matematicko-fyzikální fakulta, 2007.

Holan T., Kubon V., Oliva K., Plátek M., « On complexity of word order », *TAL. Traitement automatique des langues*, vol. 41, n° 1, p. 273-300, 2000.

Kahane S., Gerdes K., *Syntaxe théorique et formelle, Volume 1 : Modélisation, unités, structures*, Language Science Press, 2020. A paraître.

Kahane S., Yan C., « Advantages of the flux-based interpretation of dependency length minimization », *First international conference on Quantitative Syntax (Quasy)*, 2019.

Kuhlmann M., Nivre J., « Mildly non-projective dependency structures », *Proceedings of the COLING/ACL on Main conference poster sessions*, Association for Computational Linguistics, p. 507-514, 2006.

Kuhlmann M., Nivre J., « Transition-based techniques for non-projective dependency parsing », *Northern European Journal of Language Technology (NEJLT)*, vol. 2, n° 1, p. 1-19, 2010.

Lecerf Y., Ihm P., Eléments pour une grammaire générale des langues projectives, Rapport GRISA n° 1, Euratom, 1960.

Liu H., « Dependency distance as a metric of language comprehension difficulty », *Journal of Cognitive Science*, vol. 9, n° 2, p. 159-191, 2008.

Mambrini F., Passarotti M., « Non-projectivity in the Ancient Greek dependency treebank », *Proceedings of the second international conference on dependency linguistics (Depling 2013)*, p. 177-186, 2013.

Marcus S., « Sur la notion de projectivité », *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, vol. 11, p. 181-192, 1965.

McDonald R., Pereira F., Ribarov K., Hajič J., « Non-projective dependency parsing using spanning tree algorithms », *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 523-530, 2005.

McDonald R. T., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K. B., Petrov S., Zhang H., Täckström O., Bedini C., Castelló N. B., Lee J., « Universal Dependency Annotation for Multilingual Parsing. », *ACL (2)*, ACL, p. 92-97, 2013.

Mel'cuk I. A. *et al.*, *Dependency syntax : theory and practice*, SUNY press, 1988.

Miletic A., Urieli A., « Non-projectivity in Serbian : Analysis of formal and linguistic properties », *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, p. 135-144, 2017.

Nivre J., « Constraints on non-projective dependency parsing », *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Nivre J., « Non-Projective Dependency Parsing in Expected Linear Time », *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, Suntec, Singapore, p. 351-359, August, 2009.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N. *et al.*, « Universal dependencies v1 : A multilingual treebank collection », *Proceedings of LREC 2016*, p. 1659-1666, 2016.

Straka M., Hajic J., Straková J., Hajic Jr J., « Parsing universal dependency treebanks using neural networks and search-based oracle », *International Workshop on Treebanks and Linguistic Theories (TLT14)*, p. 208-220, 2015.

Tapanainen P., Jarvinen T., « A non-projective dependency parser », *Fifth Conference on Applied Natural Language Processing*, Association for Computational Linguistics, Washington, DC, USA, p. 64-71, 1997.