Évaluation linguistique de résultats de parsing syntaxique dépendanciel automatique (spaCy) d'un corpus de poèmes en prose de Pierre Reverdy

Auteur : Beckers, Xavier
Promoteur(s) : Mazziotta, Nicolas
Faculté : Faculté de Philosophie et Lettres
Diplôme : Master en langues et lettres françaises et romanes, orientation générale, à finalité approfondie
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/14775

ÉVALUATION LINGUISTIQUE DE RÉSULTATS DE PARSING SYNTAXIQUE DÉPENDANCIEL AUTOMATIQUE (*SPACY*) D'UN CORPUS DE POÈMES EN PROSE DE PIERRE REVERDY

MÉMOIRE DE MASTER

BECKERS Xavier

Master 2 en langues et lettres françaises et romanes, à finalité approfondie

Année académique 2021 – 2022

Promoteur :
Nicolas MAZZIOTTA,
Professeur

Je souhaite remercier

mes proches,
sans lesquels tout ceci n’aurait pas été possible ;

mes amis,
pour toutes ces belles années, à l’université et à distance ;

mon promoteur, le professeur N. Mazziotta,
pour sa disponibilité, sa rigueur et son expertise ;

la faculté de Philosophie et Lettres,
pour m’avoir ouvert l’esprit.

1 Introduction

Ce mémoire de linguistique appliquée propose une recherche concernant les résultats d'une analyse syntaxique automatique sur un corpus littéraire. Depuis Alan Turing (1950) et son « jeu de l'imitation », la question de la maîtrise du langage humain par les machines occupe une place particulière dans la sphère des recherches en informatique et en linguistique appliquée. Le chapeau disciplinaire du TAL (*Traitement Automatique des Langues*) englobe une série de pratiques, de la correction orthographique à la synthèse et reconnaissance vocale, en passant par l'analyse de la similarité sémantique des textes, entre lesquelles nous aimerions pointer le cas de l'analyse syntaxique automatique (ci-dessous *parsing syntaxique*). Cette dernière est en pleine mutation et connaît une évolution sensible depuis le développement des approches basées sur la syntaxe dépendancielle et l'introduction, en 2017 par une cohorte de chercheurs affiliés à Google, d'une nouvelle architecture de réseau de neurones servant à l'intelligence artificielle nommée *Transformer* (Vaswani *et al.*, 2017). Diverses bibliothèques logicielles de TAL, telles *NLTK*, *Stanford CoreNLP* ou *spaCy*, permettent aujourd'hui de mettre en place ces nouvelles technologies de parsing syntaxique souvent sous la forme de l'inclusion, dans une chaîne de traitement, d'un *parser* (ou analyseur automatique) syntaxique.

Ce travail a pour objet l'évaluation, sur un **corpus littéraire** constitué de poèmes en prose de Pierre Reverdy issus du recueil *Flaques de verre* (1929), des résultats d'un **parsing syntaxique** selon une méthode qualitative et quali-quantitative s'intéressant aux erreurs suivant **l'épistémologie de la linguistique**. En effet, ce recueil de poèmes en prose écrit en 1929 par le poète Pierre Reverdy (1889-1960), par son insertion dans la *modernité*, nous semble présenter des difficultés syntaxiques et sémantiques particulièrement intéressantes, comme l'usage spectaculaire du tiret cadratin (HE-7[1]) ou l'ellipse régulière du verbe principal (CS-1). Il est peu probable que le parser ait pu rencontrer, durant son entraînement, des particularités syntaxiques de ce type ; il s'agit

[1] Afin de nous référer aisément à des phrases précises de notre corpus, nous avons décidé d'adopter une notation combinant, en lettres, l'indicatif du poème et, en chiffres, le numéro de la phrase concernée à l'intérieur de ce poème. Ainsi, HE-7 correspond à la septième phrase du poème *L'Homme aux étoiles*. La liste exhaustive des indicatifs des poèmes est disponible en annexe (A.2).

donc d'un corpus idéal pour évaluer son adaptabilité ainsi qu'isoler des caractéristiques du *style syntaxique* de Reverdy dans ce recueil.

> (HE-7) Les arbres du jardin fermé sont sur la grille — les pointes du signal à côté de la mer — les deux battants ouverts sur l'horizon qui grince — le jour lâché — s'évade et piétine les ombres — les hommes — les étoiles tombées sur le revers.
> (CS-1) À pied, courant dans le désert des vents brouillés par le plus sombre caractère, toutes lumières inclinées vers la nuit, les têtes dépouillées des fleurs sentimentales et des sources d'esprit.

Dans un premier temps, nous introduisons la syntaxe dépendancielle, les outils contemporains d'intelligence artificielle adaptés au TAL ainsi que le choix du corpus et le fonctionnement de notre évaluation (→ 2). Nous présentons ensuite les résultats de notre évaluation, d'abord selon une méthode automatique classique (→ 3), puis en proposant une typologie des erreurs selon des critères linguistiques (→ 4). Nous terminons l'exposé par une brève présentation de l'intérêt de la littérature pour la grammaire ainsi que les possibilités offertes, pour l'analyse stylistique, par le parsing syntaxique automatique en proposant une analyse du *squelette phrastique* et des *ruptures d'équivalence* dans *Flaques de verre* (→ 5).

Les objectifs de cette recherche se centrent à trois niveaux : l'outil, les matériaux et la théorie propre à la linguistique. Tout d'abord, nous proposons ainsi une évaluation de parser selon une méthode plutôt qualitative en rupture avec les évaluations classiques sur corpus annoté selon un standard. Le parser est alors évalué selon des paramètres linguistiques permettant de mieux cerner les erreurs fréquentes, ce qui peut permettre de proposer des solutions de peaufinage[2] du modèle grâce à une meilleure compréhension du contexte d'apparition de ces erreurs. Ensuite, cette analyse est susceptible de montrer des spécificités syntaxiques de notre objet tout en exposant, par l'analyse des erreurs, à quel point ces structures divergent des structures rencontrées dans le corpus d'entraînement qui est constitué de textes non littéraires. Enfin, ce travail permet de s'interroger sur les limites linguistiques de l'analyse syntaxique issues de phénomènes comme l'ambiguïté ou des difficultés théoriques comme la représentation en dépendance de la coordination. Puisque ces limites sont responsables d'une partie des erreurs du parsing, elles sont évidemment à prendre en compte.

[2] *Fine tuning*, ajustement de faible ampleur d'un modèle statistique.

2 Préalable théorique et méthodologique

Nous présentons dans cette section les différents éléments théoriques et méthodologiques centraux pour notre travail. Le premier de ces éléments est la syntaxe dépendancielle (→ 2.1), conception linguistique selon laquelle la syntaxe de la phrase est analysable comme un ensemble de relations hiérarchiques typées entre des mots. La plupart des analyses syntaxiques en traitement automatique des langues sont aujourd'hui basées sur cette approche. Nous présentons ensuite le fonctionnement interne des modèles statistiques d'intelligence artificielle utilisés en TAL, en particulier les révolutions récentes du *transformer* et de *BERT*, afin de comprendre quelles sont les limites du travail effectué sur notre parser (→ 2.2). Enfin, nous présentons différentes méthodes d'évaluation des outils de traitement automatique des langues et argumentons en faveur de l'évaluation manuelle et centrée sur l'aspect linguistique présentée dans ce travail, dont nous expliquons la méthodologie (→ 2.3).

2.1 La syntaxe dépendancielle

La syntaxe dépendancielle est une approche de la langue s'attachant à décrire cette dernière selon des relations hiérarchiques typées (dites « de dépendance »). Elle constitue une alternative à une syntaxe des constituants comme l'analyse en constituants immédiats (ACI).

Cette approche dépendancielle connaît des développements importants depuis quelques décennies notamment, selon Imrényi et Mazziotta (2020 : 2), grâce au développement de larges *treebanks* (banques de données de phrases annotées morpho-syntaxiquement) et aux résultats obtenus dans le traitement automatique des langues.

L'origine de cette description est souvent ramenée à Lucien Tesnière et ses *Éléments de syntaxe structurale* parus en 1959, ouvrage fondateur puisque Tesnière tente de systématiser toute la grammaire des langues selon cette notion de dépendance. Cependant, il est nécessaire de nuancer cela : comme l'expliquent Mazziotta et Kahane (à paraître) ainsi que Imrényi et Mazziotta, Tesnière a eu connaissance de pratiques plus anciennes, comme les diagrammes de Clark ou Reed-Kellogg. Mazziotta et Kahane expliquent l'apport fondamental de Tesnière :

> On lui reconnaît notamment la paternité conceptuelle de la notion de *dépendance*, même si Tesnière n'emploie pas exactement ce mot. La dépendance doit être comprise ici comme la hiérarchie entre les mots qui constituent les constructions

> syntaxiques. […] Cette manière de concevoir les rapports entre les unités a deux conséquences directes. Premièrement, les relations sont elles-mêmes considérées comme des unités en plus des mots qu'elles unissent. […] Deuxièmement, l'≪ ordre structural ≫, l'ordre caché de l'organisation de ces dépendances, est différent de l'ordre linéaire, qui est quant à lui apparent. (Mazziotta et Kahane, à paraître : 2)

Nous présentons d'abord succinctement la « syntaxe structurale » de Lucien Tesnière et les caractéristiques formelles essentielles de la grammaire de dépendance avant d'évoquer un de ses prolongements qui est central dans notre travail : *Universal Dependencies*.

2.1.1 La syntaxe structurale de Lucien Tesnière

Comme mentionné *supra*, l'apport fondamental est la notion de dépendance (même si Tesnière lui-même parle plutôt de « connexions »), autour de laquelle s'articulent toutes les autres.

Dans les *Éléments de syntaxe structurale* de 1959, Lucien Tesnière commence par exposer la différence qu'il y a entre les mots isolés du dictionnaire et les mots dans une phrase, qui deviennent un ensemble organisé que l'esprit doit saisir sans que les connexions ne soient explicites :

> La phrase est un **ensemble organisé** dont les éléments constituants sont les **mots**.
>
> Tout mot qui fait partie d'une phrase cesse par lui-même d'être isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des **connexions**, dont l'ensemble forme la charpente de la phrase.
>
> Ces connexions ne sont indiquées par rien. Mais il est indispensable qu'elles soient aperçues par l'esprit, sans quoi la phrase ne serait pas intelligible. Quand je dis : *Alfred parle* (v. St. 1), je n'entends pas dire d'une part qu' « il y a un homme qui s'appelle Alfred » et d'autre part que « quelqu'un parle », mais j'entends dire tout à la fois que « Alfred fait l'action de parler » et que « celui qui parle est Alfred ». (Tesnière, 1959 : 11)

La phrase *Alfred parle* est en réalité composée de trois éléments : *Alfred, parle* ainsi que la connexion qui les unit :

> Il résulte de ce qui précède qu'une phrase du type *Alfred parle* n'est pas composée de **deux** éléments 1° *Alfred*, 2° *parle*, mais bien de **trois** éléments, 1° *Alfred*, 2° *parle* et 3° la connexion qui les unit et sans laquelle il n'y aurait pas de phrase. Dire qu'une phrase du type *Alfred parle* ne comporte que deux éléments, c'est l'analyser d'une façon superficielle, purement morphologique, et en négliger l'essentiel, qui est le lien syntaxique. (Tesnière, 1959 : 11-12)

C'est alors la *connexion* qui permet d'exprimer la pensée et donne à la phrase son caractère. Elle est donc à la base de la syntaxe structurale :

> La connexion est **indispensable** à l'expression de la pensée. Sans la connexion, nous ne saurions exprimer aucune pensée continue et nous ne pourrions qu'énoncer une succession d'images et d'idées isolées les unes des autres et sans lien entre elles.
>
> C'est donc la connexion qui donne à la phrase son caractère **organique** et **vivant**, et qui en est comme le **principe vital**.
>
> Construire une phrase, c'est mettre la vie dans une masse amorphe de mots en **établissant** entre eux un **ensemble** de **connexions**.
>
> Inversement, comprendre une phrase, c'est **saisir l'ensemble des connexions** qui en unissent les différents mots.
>
> La notion de connexion est ainsi à la **base** de toute la syntaxe structurale. On ne saurait donc trop insister sur son importance. (Tesnière, 1959 : 12)

Nous pouvons donc comprendre que la description d'une phrase, selon la syntaxe structurale, consiste en sa division en un ensemble de mots et la caractérisation de l'ensemble exhaustif des connexions entre ceux-ci. Ces connexions établissent des rapports de *dépendance* et chaque mot ne peut dépendre que d'un seul autre :

> Les connexions structurales établissent entre les mots des rapports de **dépendance**. Chaque connexion unit en principe un terme **supérieur** à un terme **inférieur**.
>
> Le terme supérieur reçoit le nom de **régissant**. Le terme inférieur reçoit le nom de **subordonné**. Ainsi dans la phrase *Alfred parle* (v. St. 1), *parle* est le régissant et *Alfred* le subordonné. […] L'ensemble des mots d'une phrase constitue donc une véritable **hiérarchie**. En principe, un subordonné ne peut dépendre que d'un seul régissant. Au contraire, un régissant peut commander plusieurs subordonnés. (Tesnière, 1959 : 13-14)

Puisque cette structure ne peut être cyclique, c'est-à-dire qu'il ne peut exister aucun chemin dans le graphe qui visite plus d'une fois un nœud, il est nécessaire qu'un nœud ne soit le subordonné d'aucun autre : Tesnière l'appelle le *nœud des nœuds* — nous l'appelons *racine* — et il s'agit, dans le cas d'une phrase verbale, du verbe principal :

> Le nœud des nœuds est généralement un nœud verbal, ainsi qu'il ressort des exemples cités jusqu'ici. Mais rien n'empêche qu'une phrase ait pour nœud central un nœud substantival, adjectival ou adverbial. Le cas est surtout fréquent dans la conversation courante et dans les titres d'ouvrages littéraires. (Tesnière, 1959 : 15)

Tesnière (1959) effectue une distinction importante dans la terminologie de notre travail : celle entre actants et circonstants. Les actants sont des compléments essentiels avec lesquels le verbe fini entretient des relations actancielles. Il en existe trois types : le sujet, le complément d'objet direct et le complément d'objet indirect. Les circonstants sont des compléments de phrase qui dépendent du verbe principal par une connexion circonstancielle.

Enfin, Tesnière établit un dernier type de lien afin de permettre la coordination (qui est extrêmement difficile à formaliser dans un paradigme de dépendance [→ 4.2.1]) : la jonction. Il s'agit d'une ligne horizontale, contrairement aux connexions qui ont toujours une orientation verticale afin de marquer cette asymétrie régissant/subordonné qui n'est pas présente dans le cas de la coordination. Selon Tesnière (1959 : 323), « la jonction consiste à ajouter entre eux des nœuds de même nature, de telle sorte que la phrase, grossie de ces nouveaux éléments, gagne en ampleur et devient par là plus longue ».

2.1.2 Caractéristiques formelles des grammaires de dépendance

La conception de Tesnière, dont les continuateurs comme Mel'čuk (1988) conservent principalement la connexion (ou dépendance), implique une série de traits formels essentiels à toute analyse dépendancielle. Dans la préface de leur ouvrage, Polguère et Mel'čuk (2009 : xiii-xv) indiquent quatre aspects primordiaux de la description en dépendance d'une phrase (ainsi que leurs conséquences formelles) : 1. la connexité de la structure syntaxique (« connectedness of the syntactic structure ») ; 2. l'orientation des relations syntaxiques (« directedness of syntactic relations ») ; 3. la relation hiérarchique stricte (« strict hierarchical organization of the syntactic structure ») ; 4. l'expressivité des relations (« 'meaningfulness' of syntactic relations »).

a) Connexité de la structure syntaxique

La connexité de la structure syntaxique indique qu'une structure syntaxique doit être connexe[3], au sens mathématique du terme. En conséquence, toute unité syntaxique faisant partie d'une phrase doit être liée syntaxiquement à une autre unité et aucune unité de la phrase ne peut être en dehors de cette structure. De même, la structure ne peut pas se ramener à deux ensembles de connexions entre lesquels aucune relation n'existe, sans quoi il s'agit de deux phrases distinctes. L'objet formel qui en résulte est un *graphe connexe* (*connected graph*, Bretto *et al.*, 2012 : 8). Il s'agit d'ailleurs de la propriété qu'utilise notre parser pour délimiter les phrases, puisqu'un graphe syntaxique qui n'est pas connexe représente plusieurs phrases (autant de phrases qu'il existe de sous-graphes connexes) ; nous y revenons dans la section adéquate (→ 2.2.2).

[3] C'est-à-dire que, dans un graphe, il existe un chemin (une *chaîne* [Bretta *et al.*, 2012 : 7]) reliant toute paire de sommets appartenant à ce graphe.

b) Orientation des relations syntaxiques

L'orientation des relations syntaxiques indique que les connexions doivent être des relations de dépendance qui sont orientées dans l'un ou l'autre sens. L'un des deux éléments est alors nommé « gouverneur ». Selon Mel′čuk, cela reflète la nature asymétrique des phrases : un composant domine les autres. Cela est montré, selon lui, par le fait que la phrase soit capable d'intégrer une nouvelle unité lexicale seulement par l'intermédiaire d'un de ses composants ; ainsi, une phrase se comporte comme son composant dominant. L'objet formel est donc un graphe connexe *orienté*[4] (*directed connected graph*), ou *digraphe connexe*.

c) Relation hiérarchique stricte

La relation hiérarchique stricte (« unicité du gouverneur syntaxique » [Mel'čul, 1988 : 24]) implique que chaque unité lexicale a toujours un et un seul gouverneur syntaxique, à l'exception d'une unité qui ne doit pas avoir de gouverneur du tout — car les dépendances ne sont pas cycliques[5]. Cette unité qui n'est pas gouvernée est appelée nœud du haut (*top node*) et est la tête de la phrase ; il s'agit du gouverneur absolu. Beaucoup de paradigmes la nomment racine (*root*). Cette tête est nécessairement unique. Cela a pour conséquence formelle que la structure syntaxique est un digraphe connexe *acyclique* (Bretta *et al.*, 2012 : 39). Cette structure est également appelée *arbre enraciné*, au sens mathématique du terme.

d) Expressivité des relations

Enfin, la propriété d'expressivité indique qu'il n'est pas suffisant de relever et orienter les relations syntaxiques entre des unités afin de les qualifier pleinement puisque cela mènerait à des confusions en ce qui concerne le rôle des unités : dans *Mother sent **Mary** to the doctor* et *Mother sent **Mary** 200 $* (Polguère et Mel'čuk, 2009 : xv), il est nécessaire d'exprimer quelle est la relation entre *sent* et *Mary*, qui est, dans la première phrase, complément d'objet direct, et dans la seconde, complément d'objet indirect. En effet, selon les auteurs une relation syntaxique porte bien plus d'information linguistique que simplement l'indication de l'organisation hiérarchique

[4] C'est-à-dire que les arêtes du graphe ne peuvent être empruntées que dans un sens.

[5] Un cycle, en théorie des graphes, est une succession finie d'arêtes ne contenant pas deux fois une même arête et dont les deux extrémités coïncident (Bretta *et al.*, 2012 : 7-8). Lorsqu'un graphe est cyclique, il est donc possible de visiter deux fois un même sommet sans jamais se déplacer plusieurs fois le long de la même arête. Un graphe ne contenant pas de cycle est dit *acyclique*.

de la phrase, mais est un pont entre le sens de la phrase et sa forme de surface. Une relation syntaxique ne correspond cependant pas à un sens spécifique : le sujet syntaxique peut être agent, patient ou locatif par exemple. Les auteurs le résument : « *syntactic relations do correspond to semantic roles (and vice versa) but these correspondences are by no means direct or systematic* » (Polguère et Mel'čuk, 2009 : xv).

De nombreux formalismes respectant ces quatre propriétés ont été présentés, parmi lesquels *Universal Dependencies*. La grande quantité de banques de phrases annotées selon ses règles a conduit naturellement, grâce aux données d'entraînement disponibles, de nombreux outils de TAL à baser leur analyseur de syntaxe sur UD.

2.1.3 Universal Dependencies (UD)

UD est donc un prolongement des théories antérieures de la grammaire dépendancielle, en particulier la Théorie Sens-Texte de Mel'čuk (1988), qui est apparu afin de soutenir la création d'immenses corpus adaptés au traitement automatique des langues. Selon ses contributeurs principaux (de Marneffe *et al.*, 2021 : 255 ; Nivre *et al.*, 2020 : 4034), UD n'est pas qu'un ensemble de règles permettant l'annotation morphosyntaxique homogène entre les langues, mais également une communauté et une collection de corpus :

> Universal dependencies (UD) is at the same time a framework for crosslinguistically consistent morphosyntactic annotation, an open community effort to create morphosyntactically annotated corpora for many languages, and a steadily growing collection of such corpora. In all these respects, UD has undeniably been very successful, growing in only six years from ten treebanks and a dozen researchers to 183 treebanks for 104 languages with contributions from 416 researchers around the world. UD treebanks are now widely used in natural language processing research, including but not limited to research on syntactic and semantic parsing, and increasingly also in linguistic research, particularly on psycholinguistics and word order typology. […] The goal of UD is to offer a linguistic representation that is useful for morphosyntactic research, semantic interpretation, and for practical natural language processing across different human languages. It therefore puts an emphasis on simple surface representations that allow parallelism between similar constructions across different languages, despite differences of word order, morphology, and the presence or absence of function words. (De Marneffe *et al.*, 2021 : 255-256)

Le formalisme UD est donc lui aussi basé sur une perspective de grammaire de dépendance : un élément a un *gouverneur* (*head*) et éventuellement des *dépendants* (*dependents*). Les mots sont donc organisés dans une structure en arbre avec le prédicat

principal comme racine (De Marneffe, 2021 : 257). Il est important de noter que la notion de mot morphosyntaxique de UD ne correspond pas toujours avec les unités orthographiques ou phonologiques : les clitiques doivent être séparés et traités comme mots indépendants (an. *'s*) ; les mots composés doivent recevoir un traitement spécial, car certaines langues intègrent des marques comme des espaces (fr. *pomme de terre*) (De Marneffe, 2021 : 259). De même, les éléments des unités agglutinées sont séparés[6], comme fr. *du* qui est analysé comme la contraction de *de* et *le*. On parle, dans ces derniers cas, de *multiword token*, token à mots multiples, car un token orthographique correspond à plusieurs mots syntaxiques (Nivre *et al.*, 2020 : 4035). Ces auteurs ajoutent que UD est basé sur un point de vue lexical de la syntaxe, ce qui exclut toute tentative de diviser les mots en morphèmes sans pour autant se limiter à la définition orthographique du mot.

En plus de la grammaire de dépendance, UD définit un ensemble d'étiquettes de parties du discours (nommées UPOS, *Universal Part Of Speech*) ainsi que des traits morphologiques (*features* : genre, temps, cas, degré, aspect) et leur liste de valeurs possibles. Il s'agit donc d'un environnement permettant une qualification avancée et précise de phrases de toutes les langues du monde, rendant possible la comparaison. Nous n'utilisons cependant pas les informations UPOS et de morphologie dans le cadre de ce travail, car elles ne concernent pas le parser syntaxique (→ 2.2.2).

Pour la description syntaxique des langues, UD a développé un ensemble de 37 relations grammaticales universelles que nous détaillons à la fin de cette section. L'ensemble est fermé, mais un mécanisme de sous-typage est possible grâce à l'usage des deux-points permettant l'étiquetage de constructions spécifiques à la langue étudiée :

> Perhaps the most distinctive feature of UD is its taxonomy of grammatical relations between words. Each dependent of a head, and also any function words that belong with a head, are connected to the head via a grammatical relation drawn from a universal typology of 37 grammatical relations. […] The *root* relation is used for the root of the sentence, with a dummy head that does not need to be explicit. The *dep* relation is used when no other relations are deemed appropriate. […] The set of

[6] Notons que notre parser ne sépare pas les agglutinations ; nous les conservons donc en l'état dans notre travail. Lorsque l'agglutination contient un mot fonctionnel (souvent une préposition), l'ensemble est considéré comme occupant la fonction de ce mot dans la phrase et le reste (souvent un déterminant) est ignoré.

> allowed relations is closed, but UD allows relation subtypes separated from the main relation by a colon to provide further distinctions or to capture language-specific constructions. For example, a number of languages mark relative clauses as *acl:relcl* and predeterminers as *det:predet*. (De Marneffe *et al.*, 2021 : 265)

Ces relations sont classées selon une typologie qui se divise en trois sortes d'unités, avec une division fondamentale entre les éléments nominaux et les éléments propositionnels, qui se rapportent à un prédicat :

> In more detail, UD assumes a simple typology of three kinds of phrasal units (which might minimally be just a single word):
>
> 1. Nominals: the primary means for referring to entities
> 2. Clauses: the primary means for referring to events
> 3. Modifiers: the canonical attributive modifiers of nominals, clauses, and other modifiers
>
> The distinction between nominals and clauses is fundamental to UD, which systematically uses different dependency relations in the two types of structures. (De Marneffe, 2021 : 257-258).

Cette division est croisée avec la distinction entre *core arguments* (arguments dits essentiels, actants chez Tesnière) du prédicat et *oblique modifiers* (arguments périphériques, circonstants chez Tesnière) :

> In classifying grammatical relations, UD distinguishes the **core arguments** of a predicate, essentially subjects and objects, from all other dependents at the clause level, collectively referred to as **oblique modifiers**. The core–oblique distinction is commonly assumed in typological linguistics (see, e.g., Thompson 1997; Andrews 2007) and is ultimately an information packaging distinction. (De Marneffe *et al.*, 2021 : 266)

Cette typologie est représentée dans la figure 0, qui reprend les 37 étiquettes de relation.

La plupart de ces étiquettes permettent de qualifier la relation d'un dépendant à son gouverneur le plus précisément possible tout en conservant un niveau de généralité élevé afin de pouvoir annoter des phrases issues de n'importe quelle langue. Nous pouvons observer par exemple l'étiquette de sujet nominal *nsubj*, de complément déterminatif *nmod* ou encore d'épithète *amod*. D'autres, comme *conj*, *list* et *parataxis* (ainsi que dans une moindre mesure, *appos*) permettent de qualifier les relations d'équivalence ou d'entassement (Rossi-Gensane, 2017, → 4.2.1), qui ne sont pas strictement des relations de dépendance, tout en conservant les propriétés

mathématiques de l'arbre syntaxique. Les étiquettes *fixed*, *flat*, *compound*, *goeswith*, *reparandum* et *punct* permettent le traitement des figements et des unités à mots multiples, des mots divisés en plusieurs parties et de la ponctuation. Enfin, l'étiquette *root* indique la racine de la phrase, *orphan* permet de traiter les cas d'ellipse sans nécessiter la création d'un nœud fantôme occupant la fonction du mot effacé et *dep* (*unspecified dependency*) est l'étiquette utilisée lorsqu'il est impossible de qualifier plus précisément une relation. Une liste contenant toutes les relations de notre corpus accompagnées d'exemples est disponible en annexe (A.4).

Typology of the syntactic relations.

Head \ Dependent	Nominals	Clauses	Modifier words	Function words
Clausal core arguments	nsubj obj iobj	csubj ccomp xcomp		
Clausal non-core arguments	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Fig. 0 — Relations syntaxiques de l'environnement UD
(De Marneffe *et al.*, 2021 : 267)

Le choix de la tête des arguments, c'est-à-dire le nœud d'un syntagme duquel dépendent tous les autres et recevant l'étiquette de relation syntaxique de l'ensemble[7], est une décision critique sujette à débat : là où un environnement concurrent comme SUD (*Surface-Syntactic Universal Dependencies*, Gerdes *et al.*, 2018) considère que la tête est « fonctionnelle » — c'est-à-dire que c'est le mot fonctionnel, la préposition ou la conjonction, qui régit l'argument — UD a décidé de sélectionner le mot lexical principal, souvent le substantif ou le verbe, comme tête de l'argument. Selon De Marneffe *et al*. (2021 : 303), cela permet un meilleur parallélisme entre les langues et

[7] Dans *Pierre mange une pomme*, la tête du syntagme *une pomme* est le mot pomme, qui dépend donc directement du verbe *mange* en tant que complément d'objet direct (*obj*). Le déterminant *une* dépend du substantif auquel il se rapporte, *pomme*, selon la relation d'actualisateur (*det*).

facilite l'extraction de l'information. Ce choix est contraire à la tradition de la grammaire dépendancielle selon Osborne et Gerdes (2019 : 2) : « *The decision to subordinate function words to content words is, as stated above, contrary to the DG* [*Dependency Grammar*] *tradition* ». Les auteurs considèrent qu'il s'agit d'une faiblesse de UD du point de vue linguistique et qu'il est préférable de privilégier les têtes fonctionnelles :

> Linguistic considerations have revealed that the current UD annotation scheme results in structures that are a mixture of semantic and syntactic motivations. These structures are hence not well-motivated from the linguistic point of view. As an alternative, we have advocated for a more traditional annotation scheme, one that consistently elevates syntactic criteria for determining headhood over semantic criteria. This alternative annotation scheme positions auxiliary verbs as heads over content verbs, the copula as head over predicative elements, and adpositions and subordinators as heads over nouns and verbs. (Osborne et Gerdes, 2019 : 24).

Nous argumentons également en faveur de la sélection des mots fonctionnels comme têtes après l'analyse des erreurs (→ 4.2.14).

Le postulat linguistique principal de UD est celui qu'il y a quelque chose de commun entre les langues humaines qui dépasse les variations superficielles des langues particulières. D'un point de vue pratique, en particulier en TAL, l'objectif est également de fournir un environnement commun permettant de construire aisément des systèmes de traitement automatique des langues en transférant efficacement des connaissances acquises sur d'autres langues afin d'approcher l'objectif d'un parser universel qui fonctionnerait pour toutes les langues (De Marneffe, 2021 : 302).

La conception se base sur la recherche d'un **compromis** entre la précision, permettant de ne pas annoter de la même façon des éléments qui sont différents, et la généralisation aux différentes langues sans pour autant obscurcir les différences réelles (De Marneffe, 2021 : 302). Les auteurs présentent une liste de critères fondamentaux de la construction d'UD :

> The secret to understanding the design and success of UD is to realize that the design is a very subtle **compromise** between a number of competing criteria:
>
> 1. UD needs to be reasonably satisfactory on linguistic analysis grounds for individual languages—a journeyman's universal grammar.
> 2. UD needs to be good for linguistic typology: It should bring out crosslinguistic parallelism across languages and language families.
> 3. UD must be suitable for rapid, consistent annotation by a human annotator.

4. UD must be easily comprehended and used by non-linguist users with prosaic needs.

5. UD must be suitable for computer parsing with high accuracy.

6. UD must support well downstream language understanding tasks, such as relation extraction, reading comprehension, machine translation, and so on. (De Marneffe *et al.*, 2021 : 302-303)

Le succès fut au rendez-vous, puisque Nivre *et al*. (2021 : 4040) recensent 157 treebanks annotés avec UD, pour 90 langues différentes en 2019 lors de la publication de la norme UD 2.5.

Puisque nous travaillons avec des banques de données d'arbres syntaxiques, autant pour considérer l'entraînement du parser que pour l'évaluer, il est important de saisir la forme sous laquelle se présentent ces treebanks : le format de fichier CoNLL-U (→ 2.1.3.1). Nous présentons ensuite le corpus UD important dans notre travail, puisqu'il s'agit du corpus sur lequel se base notre parser : UD-Sequoia (Candito et Seddah, 2012 ; → 2.1.3.2).

2.1.3.1 CoNLL-U : format des treebanks UD

Le format CoNLL-U est une variation standardisée du format CoNLL-X développée dans le cadre de la dixième *Conference on Computational Natural Language Learning* (CoNLL). C'est ce format qui est utilisé dans l'environnement UD. Dans ce format, chaque ligne (à l'exception des lignes de commentaire et des lignes vierges indiquant la division entre deux phrases) correspond à un token et présente dix champs permettant de rassembler les informations morphosyntaxiques de celui-ci. Ces champs peuvent être vides, mais ils sont alors représentés par le tiret bas. Il s'agit des champs suivants :

- ID : index du token dans la phrase, commençant à 1 pour chaque nouvelle phrase et incrémenté selon la linéarité du texte ;
- FORM : forme du *token* dans le texte ;
- LEMMA : lemme du *token* ;
- UPOS : étiquette universelle de partie du discours (*adj* [adjectif], *adv* [adverbe], *intj* [interjection]… ;
- XPOS : étiquette de partie du discours spécifique à la langue ;
- FEATS : liste d'éléments morphologiques (genre, nombre, mode…) de l'inventaire universel ou d'une extension spécifique à une langue ;

- HEAD : ID du gouverneur, du point de vue de la syntaxe dépendancielle, du mot courant ; 0 s'il s'agit de la racine de la phrase ;
- DEPREL : étiquette de la relation au gouverneur selon le schéma d'annotation UD ; *root* si HEAD vaut 0 ;
- DEPS : relations *enhanced* permettant la définition d'un graphe de dépendance enrichi et soumis à des règles moins strictes (le graphe peut ne pas être un arbre enraciné) afin de traiter des cas particuliers ;
- MISC : toute autre annotation ;

Un treebank est donc constitué d'un ou plusieurs fichiers avec l'extension *conllu*. Resnik et Lin (2010 : 277) indiquent que les corpus adaptés à l'entraînement de modèles statistiques se divisent en trois ou quatre sections séparées dans des fichiers différents : les données d'entraînement (*train*), grâce auxquelles le modèle est entraîné en réalisant des prédictions[8] et en ajustant les valeurs en fonction du résultat ; les données de développement (*dev*) permettant de contrôler et optimiser l'entraînement (notamment en évitant le surajustement [Charniak, 2018 : 79-82], une adaptation excessive du modèle aux données d'entraînement qui le préviendrait de réaliser des prédictions correctes sur des données nouvelles) ; les données de test (*test*), sur lesquelles sont réalisées les évaluations en vue d'une publication. Certains corpus intègrent un segment *devtest* dont l'objectif est l'évaluation formative. Les auteurs indiquent que les données sont généralement divisées comme suit : 70 % des données sont intégrées dans *train*, 20 % dans *dev* et 10 % dans *test*.

2.1.3.2 Le corpus Sequoia

Le corpus nommé UD-Sequoia, conversion en dépendances du corpus Sequoia (Candito et Seddah, 2012), nous intéresse particulièrement puisque c'est lui qui a été choisi par les développeurs de spaCy pour entraîner le parser *fr-dep-news-trf* qui sert d'objet à ce travail (→ 2.2.2.1). Le corpus Sequoia comprend l'analyse de 3204 phrases (69246 tokens) issues d'Europarl français, du journal *l'Est Républicain*, de Wikipédia Fr et des documents de l'Agence Européenne du Médicament (Candito et Seddah, 2012 : 1). Il s'agit donc d'un ensemble de phrases relevant plutôt du genre informatif.

[8] Le résultat final, déterministe, d'un modèle statistique est souvent nommé *prédiction* du modèle. Un modèle statistique réalise donc des prédictions concernant une tâche particulière en fonction de données d'entrée.

Ces phrases sont à l'origine annotées selon une analyse syntaxique en constituants, mais le corpus a été converti automatiquement en dépendances dans un second temps, selon les règles d'annotation de UD.

UD-Sequoia est lui aussi divisé en trois sections. *Train* est composé de 50517 tokens, *dev* 10002 et *test* 10048. La répartition des tokens est donc plutôt 70 %, 15 % et 15 % dans ce cas. Le segment *test* est intéressant pour proposer une première évaluation automatique et solide de notre parser, puisque ce dernier n'a auparavant été confronté qu'aux données *train* et *dev*. Nous faisons désormais référence aux segments de Sequoia sous la forme suivante : Sequoia-Train pour le segment d'entraînement, Sequoia-Test pour le segment d'évaluation et Sequoia-Dev pour le segment de développement.

Sequoia-Train contient 47 étiquettes de relation différentes en comptant les relations spécifiques, mais n'intègre pas les étiquettes *clf*, *compound*, *list* et *reparandum*. Puisque celles-ci sont absentes du corpus d'entraînement du parser, il est impossible que le modèle réalise une prédiction d'une de ces étiquettes et nous pouvons les mettre de côté. L'annexe 3 (A.3) explique et illustre chacune des étiquettes de UD-Sequoia ; il est pertinent pour le lecteur n'étant pas initié à ces relations de garder cette annexe à disposition pour la lecture des annotations syntaxiques présentes dans le travail, bien qu'elles soient généralement expliquées dans chaque cas.

2.2 L'intelligence artificielle au service du traitement automatique des langues

Afin de saisir pleinement les possibilités qui s'offrent à nous pour les perspectives de ce travail à partir des résultats obtenus, il est important de comprendre le fonctionnement basique des outils aujourd'hui utilisés en traitement automatique des langues : les réseaux de neurones. Nous commençons donc cette partie par une présentation du fonctionnement des réseaux de neurones (ou modèles statistiques) d'intelligence artificielle utilisés aujourd'hui (→ 2.2.1) avant de nous pencher sur l'aspect concret des chaînes de traitement utilisées en TAL, en particulier la chaîne fr-dep-news-trf développée par spaCy (→ 2.2.2).

2.2.1 Modèles statistiques en TAL

Nous présentons d'abord, afin de poser de solides bases pour la compréhension des mécanismes ultérieurs, le *perceptron*, invention de Frank Rosenblatt en 1957 qui est à

l'origine de tous les modèles utilisés encore à présent, ainsi que les *feed-forward neural networks* (FFNNs) (→ 2.2.1.1). Nous abordons ensuite l'architecture du *transformer* (→ 2.2.1.2), à la base de BERT et sa version française, CamemBERT, modèles entraînés ayant permis des avancées significatives dans le domaine du TAL. BERT est à l'origine de la plupart des outils de TAL modernes, dont les nôtres.

2.2.1.1 Perceptron, feed-forward neural network

Le premier type de modèle statistique *moderne* a été proposé en 1957 par Frank Rosenblatt (Rosenblatt, 1958). Il s'agit d'un classifieur binaire, permettant donc de séparer des objets en deux catégories distinctes, basé sur les observations des neurones humains : pour être activé, le neurone doit dépasser un certain seuil qui est atteint — ou non — selon l'intensité et le type de stimulations qu'il reçoit. Les stimulations peuvent ainsi être plus ou moins fortes, mais également soit excitatrices, soit inhibitrices (Rosenblatt, 1958). Ainsi, si la somme des stimulations dépasse un certain seuil, le neurone s'active et émet à son tour une stimulation. La terminologie de ce type d'architecture est largement inspirée de la biologie : Charniak (2018 : 3) dit que le neurone a généralement plusieurs entrées nommées dendrites, un corps et une simple sortie nommée axone.

La figure 1 montre une représentation graphique d'un neurone artificiel, généralisation du perceptron.

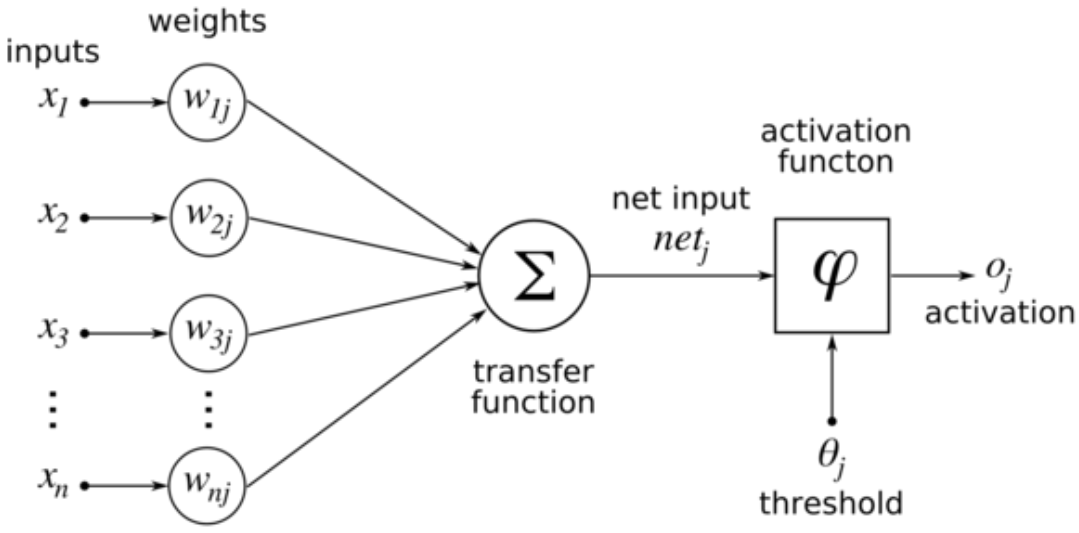


Fig. 1 — Représentation graphique d'un neurone artificiel
(Rohrmanstorfer *et al.*, 2021)

Concrètement, l'état de la cellule est donné comme ceci :

> A perceptron consists of a vector of weights $w = [w_1 \ldots w_n]$, one for each input, plus a distinguished weight b, called the bias. We call w and b the parameters of the perceptron. [Afin d'obtenir l'état de notre cellule], we multiply each perceptron input by the weight for that input and add the bias. If this value is greater than zero, we return 1, otherwise 0. Perceptrons, remember, are binary classifiers, so 1 indicates that x is a member of the class and 0, not a member. (Charniak, 2018 : 4)

Puisque le perceptron est un classifieur binaire, sa « fonction d'activation », définissant le taux d'activation du neurone et donc sa sortie, est binaire également et ne renvoie que les valeurs 1 ou 0. En pratique, pour la plupart des modèles statistiques plus complexes, la fonction d'activation renvoie des valeurs dans l'intervalle [0,1] ou [-1,1]. Lorsque plusieurs sorties sont concurrentes, la sortie considérée *prédite* est alors la sortie possédant le taux d'activation le plus élevé, donc la probabilité la plus élevée.

L'entraînement de ce type de modèle est effectué en ajustant les paramètres (*w* et *b*) de chaque neurone en commençant par les paramètres les plus proches de la sortie en avançant pas à pas pour améliorer les prédictions du modèle sur l'ensemble d'entraînement.

Cependant, un classifieur binaire n'est pas du tout suffisant dans notre cas : que ferions-nous, pour une analyse syntaxique, d'un modèle qui n'est capable que de distinguer deux classes d'éléments ? Il est alors nécessaire d'augmenter la complexité en ajoutant des neurones répartis en **couches**, ce que l'on nomme *feed-forward neural network*, ou FFNN ; les FFNNs permettent un classement non linéaire :

> In feed-forward neural networks (FFNNs), sets of neurons are organised in layers, where each neuron computes a weighted sum of its inputs. Input neurons take signals from the environment, and output neurons present signals to the environment. Neurons that are not directly connected to the environment, but which are connected to other neurons, are called hidden neurons. […] Sets of neurons organised in several layers can form multilayer, forwardconnected networks. The input and output layers are connected via at least one hidden layer, built from set(s) of hidden neurons. […] Multilayer feed-forward networks using sigmoid threshold functions are able to express non-linear decision surfaces. Any function can be closely approximated by these networks, given enough hidden units. (Staudemeyer et Morris, 2019 : 7-8)

Les modèles statistiques d'intelligence artificielle fonctionnent donc sur ces principes. Adapté au TAL, un FFNN ne serait pourtant pas efficace, car, en langue, la **notion de contexte** est particulièrement importante. Les chercheurs se sont donc mis à utiliser des réseaux de neurones permettant de « stocker en mémoire » les éléments sur

lesquels le modèle avait déjà fourni des prédictions afin de prendre en compte le contexte : les réseaux de neurones récurrents (RNNs ; Amidi et Amidi, 2019), et une de leur évolution, le LSTM (*long short-term memory* ; Charniak, 2018 : 89). Ces modèles fonctionnent sur des systèmes permettant de créer une illusion de mémoire des prédictions précédentes : les modèles effectuent des **itérations**, réalisant une prédiction à chaque fois, et l'état de la cellule de l'itération précédente devient un paramètre d'entrée de l'itération suivante. Plutôt qu'un lien direct entre sortie et entrée, le LSTM utilise un ensemble de valeurs (dont l'ensemble est appelé « mémoire ») avec lesquelles le modèle peut interagir à chaque itération, en sélectionnant lesquelles il est pertinent d'oublier, de conserver et de prédire (Olah, 2015). Il s'agit de modèles très efficaces dans le **traitement de séquences**, ce qui les rend intéressants en TAL. Cependant, en plus de montrer des difficultés dans le traitement des séquences très longues, l'inconvénient majeur des LSTM est que leur entraînement est lent : il est difficilement *parallélisable*, c'est-à-dire qu'il est difficile d'utiliser plusieurs unités de calcul simultanément pour accélérer le processus.

La révolution vient en 2017 grâce aux *transformers* et est basée sur les *mécanismes d'attention* (Vaswani *et al.*, 2017). Il s'agit de mécanismes qui consistent en la prise en compte directe des éléments du contexte. En effet, des *têtes d'attention* lient directement l'itération actuelle avec des éléments contextuels à l'aide d'opérations combinant des *requêtes* (*query*) avec des *clés* (*keys*) et des *valeurs* (*values*). Lors de chaque itération, le modèle est donc capable de *prêter attention* à son contexte en sélectionnant les valeurs utiles incluses dans celui-ci lors de chaque prédiction.

2.2.1.2 Transformers

Les transformers (fig. 2), présentés en 2017 dans l'article devenu célèbre *attention is all you need*, sont « rapidement devenus l'architecture dominante pour le traitement automatique de la langue, surpassant en performance les réseaux de neurones alternatifs […] autant dans les tâches de compréhension que de génération du langage naturel » (Wolf *et al.*, 2020 : 1 ; nous traduisons). Le modèle, basé sur l'attention — en particulier le mécanisme de *self-attention*, attention à soi-même — permet, notamment grâce à la parallélisation facilitée, des résultats surprenants pour des temps d'entraînement courts par rapport aux modèles précédents (Vaswani *et al.*, 2017 et Wolf *et al.*, 2020). Bien qu'il ait été à l'origine développé pour la traduction, Vaswani *et al.* (2017 : 1)

expliquent que le modèle « *generalizes well to other tasks by applying it successfully to English constituency parsing* ».

La structure se décompose en deux parties : l'encodeur (à gauche) et le décodeur (à droite). Dans un contexte traductif, l'encodeur correspond à la phase de prise en compte de la séquence à traduire dans son ensemble alors que le décodeur correspond à la phase de production de la séquence traduite. Les auteurs expliquent que le travail de l'encodeur est effectué une fois par séquence alors que le décodeur est exécuté pour chaque mot de la traduction l'un après l'autre.

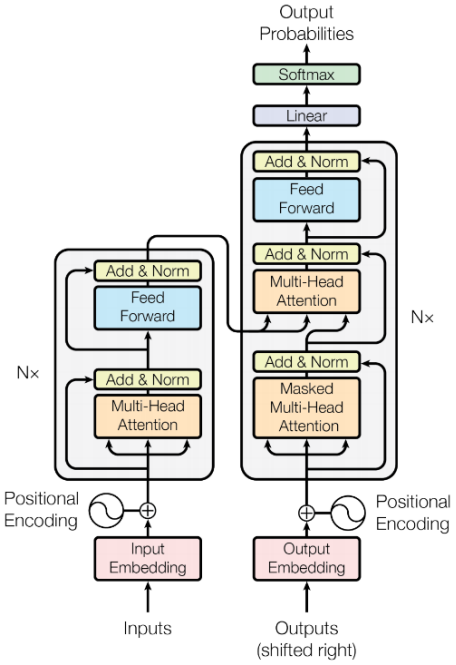


Figure 1: The Transformer - model architecture.

Fig. 2 — Architecture d'un transformer (Vaswani *et al.*, 2017 : 3)

Du côté de l'encodeur, qui nous intéresse, après une phase de vectorisation[9] de la séquence d'entrée (*input embedding*) à l'aide d'un dictionnaire prédéfini, un traitement mathématique est appliqué afin d'encoder la position de chaque unité dans la séquence (*positional encoding*). Les données entrent ensuite dans la première des N couches constituées d'un mécanisme d'attention (*Multi-Head*[10] *Attention*) puis d'un FFNN (Vaswani *et al.*, 2017 : 3). C'est ce mécanisme d'attention qui est révolutionnaire dans ce cas : il s'agit d'un mécanisme de *self-attention*, dans lequel la séquence construit

[9] Représentation de l'entrée sous forme d'un vecteur de valeurs.

[10] « Multi-Head Attention consists of several attention layers running in parallel » (Vaswani *et al.*, 2017 : 4).

elle-même des relations d'attention et sélectionne donc, pour chaque unité, différentes valeurs d'unités environnantes importantes pour leur caractérisation personnelle. L'issue de cet encodeur est ensuite utilisée, pour la traduction, dans un mécanisme d'attention du décodeur, mais cela dépasse notre propos.

2.2.1.3 BERT et CamemBERT

C'est à partir du concept de transformer qu'une équipe rattachée à Google a mis au point en 2019 un modèle ainsi qu'une stratégie de préentraînement capable de construire une *représentation de la langue* (Devlin *et al.*, 2019 : 4171), pouvant ensuite être adapté à diverses tâches de traitement automatique de la langue. Ce développement, toujours de premier plan aujourd'hui, a été nommé BERT (*Bidirectional Encoder Représentations from Transformers* [Devlin *et al.*, 2019]). Une équipe de Facebook a ensuite proposé la variante RoBERTa (*Robustly Optimized BERT pre-training Approach* [Liu *et al.*, 2019]), de laquelle est inspirée la première version française monolingue de BERT, CamemBERT (Martin *et al.*, 2020). Enfin, c'est sur CamemBERT qu'est basée la chaîne de traitement de spaCy que nous avons sélectionnée (→ 2.2.2.1).

BERT consiste en une pile d'encodeurs de transformers. Il existe en deux versions, $\text{BERT}_{\text{BASE}}$ et $\text{BERT}_{\text{LARGE}}$, le premier contenant 12 encodeurs pour 110 millions de paramètres et le second 24 encodeurs pour 340 millions de paramètres (Devlin *et al.*, 2019 : 4173). L'intérêt vient de l'entraînement en deux étapes, comme représenté à la figure 3 :

> There are two steps in our framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For finetuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters. […] For the pre-training corpus we use the BooksCorpus (800M words) and English Wikipedia (2,500M words). […] Fine-tuning is straightforward since the selfattention mechanism in the Transformer allows BERT to model many downstream tasks— whether they involve single text or text pairs—by swapping out the appropriate inputs and outputs. […] For each task, we simply plug in the taskspecific inputs and outputs into BERT and finetune all the parameters end-to-end. (Devlin *et al.*, 2019 : 4173-4175)

BERT est donc préentraîné à l'aide de données non annotées. À chaque itération d'entraînement, l'algorithme présente deux phrases dont 15 % des mots ont été

masqués[11] et demande au modèle de prédire 1. si la phrase A précède la phrase B ; 2. quelle était la séquence originelle ? Cela permet ensuite, en peaufinant[12] le modèle pour des tâches spécifiques (en substituant des sorties adaptées au problème aux sorties originelles[13]), d'obtenir des résultats à la pointe de l'état de l'art dans un grand nombre de tâches[14] avec des moyens limités. En effet, là où le préentraînement demande quatre jours sur des configurations matérielles très importantes, le peaufinage est réalisable en à peine quelques heures à l'aide d'un ordinateur personnel.

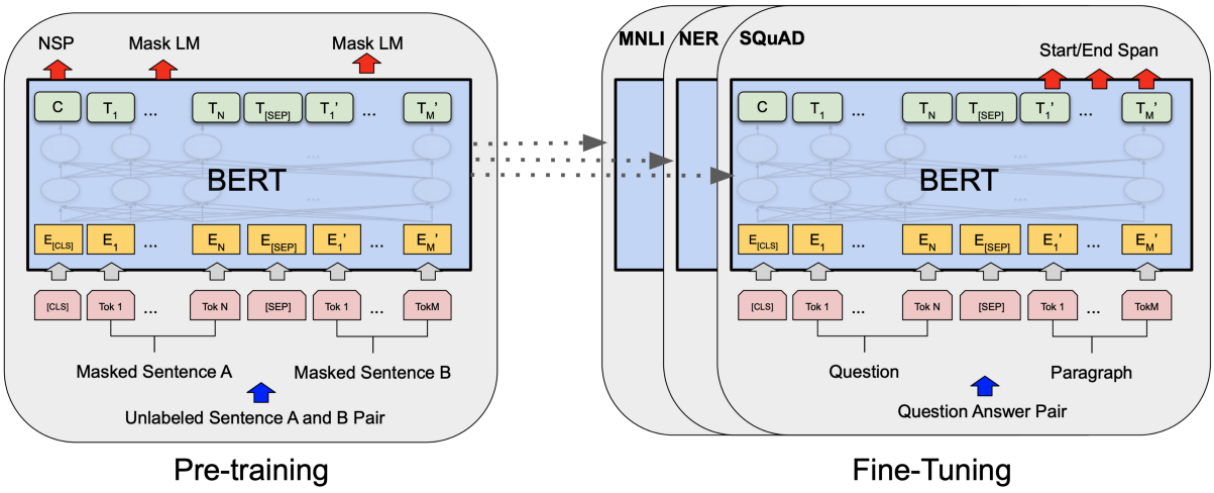


Fig. 3 — préentraînement et peaufinage de BERT

CamemBERT (Martin *et al.*, 2020) fonctionne sur le même principe et améliore les repères dans différentes tâches de TAL pour le français, entre autres l'analyse dépendancielle.

Il est particulièrement important de retenir que les informations que reçoit le modèle est uniquement une matrice constituée des vecteurs des mots en entrée, constitués eux-mêmes de la vectorisation[15] du mot réalisée auparavant par un autre réseau de neurones et stockée dans un dictionnaire, de sa position et de son appartenance à la phrase A ou la

[11] Parmi ces 15 %, 80 % sont remplacés par un token [MASK], 10 % par un mot aléatoire et 10 % par le mot original.

[12] Le *fine-tuning*, ou peaufinage, consiste en un ajustement de faible ampleur des valeurs du modèle postérieur à l'entraînement principal. Il est généralement effectué grâce à des ensembles de données plus réduits, mais contenant des données précises adaptées à un problème donné.

[13] « pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks » (Devlin *et al.*, 2019 : 4171)

[14] « BERT advances the state of the art for eleven NLP tasks » (Devlin *et al.*, 2019 : 4172).

[15] Mécanisme transformant un élément en un ensemble de valeurs décrivant ses traits définitoires. Dans le cas de mots, il s'agit par exemple d'informations distributionnelles ; cette vectorisation est réalisée par un réseau de neurones produisant de 128 à 256 valeurs.

phrase B. Le modèle ne reçoit en effet aucune information encodée par l'utilisateur comme des informations grammaticales d'appartenance à une classe particulière. Cela permet de réduire grandement l'action humaine et améliore la transférabilité du modèle à d'autres langues ou d'autres tâches.

2.2.2 Chaînes de traitement, fonctions et formats : le cas de spaCy

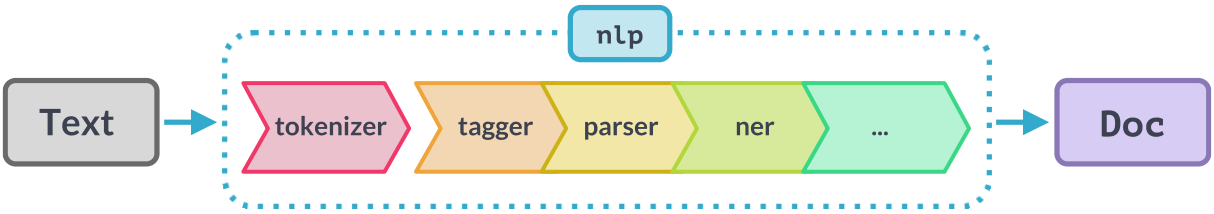


Fig. 4 — Chaîne de traitement générique de spaCy[16]

SpaCy est une bibliothèque logicielle *open source* développée par Explosion, une entreprise spécialisée dans le développement d'outils d'intelligence artificielle et TAL pour développeurs. Sa première version est introduite en 2015 et elle est toujours mise à jour aujourd'hui, avec la version 3.0 déployée le 1er février 2021 et les sous-versions 3.1 et 3.2 qui ont suivi. SpaCy[17] inclut des fonctions de division du texte en tokens, de **parsing syntaxique dépendanciel**, de reconnaissance des entités nommées (objets, personnes, entreprises…), d'évaluation de la similarité d'énoncés…

La bibliothèque permet de mettre en place des chaînes de traitement qui prennent en entrée un texte et, après son traitement par différents modules, retourne un document contenant les annotations linguistiques recherchées dans un format structuré.

L'objectif des chaînes de traitement dans notre cas est de réaliser des prédictions sur un matériau linguistique, c'est-à-dire, pour rappel, de proposer des annotations à l'aide de moyens déterministes sur un matériau inconnu à l'aide de données préalables définissant des routines généralisables. Ces prédictions sont réalisées soit à l'aide de ressources externes développées par des opérateurs humains (entrées lexicales avec leurs attributs hors contexte, règles de tokenisation, règles de lemmatisation, représentations multidimensionnelles de mots permettant de définir un degré de similarité entre eux…) généralement spécifiques à une langue en particulier, soit à

[16] *Text* représente les données d'entrée sous forme textuelle ; *Doc* représente les données formatées de sortie.

[17] https://spacy.io/usage/spacy-101.

l'aide de modèles statistiques d'intelligence artificielle — du type de ceux que nous avons présentés ci-dessus — entraînés à l'aide de corpus préparés. Certains modules utilisent évidemment des ressources des deux types.

Les chaînes de traitement disponibles nativement avec spaCy ont une forme du type représenté à la figure 4. La chaîne de traitement en particulier ([encadré bleu *nlp*]) contient différents éléments avec un aspect modulable : à l'exception notable de certains, comme le *tokenizer* (division du texte en tokens) qui est nécessairement une étape préalable, les composants ne partagent pas toujours leurs données et les résultats de l'un ne conditionnent pas ceux de l'autre[18] ; par exemple, le *ner* (*named-entity recognizer*), module s'occupant de prédire les entités nommées, ne nécessite pas que les données soient déjà annotées syntaxiquement, mais seulement que le texte ait déjà été divisé en tokens et il pourrait donc être absent ou positionné plus tôt dans la chaîne sans que cela n'influence les performances du parser syntaxique.

Le *tokenizer* est le module de tokenisation. Son rôle est de diviser le texte en tokens utilisables par les autres éléments de la chaîne. Après avoir divisé le texte selon ses espaces, le module procède récursivement de gauche à droite en appliquant deux règles : **1.** est-ce que la sous-chaîne de caractères correspond à une règle d'exception du *tokenizer* (par exemple, en français, les déterminants élidés ne sont pas suivis d'une espace, mais doivent pourtant être traités séparément) ? **2.** est-ce que la sous-chaîne contient un ou plusieurs caractères au début ou à la fin qui peuvent être séparés (comme les marques de ponctuation, particulièrement la virgule en français) ? Ces deux règles dépendant de la langue concernée, le *tokenizer* fait appel à des ressources externes sous la forme de listes de règles de tokenisation et d'exceptions.

Entrée : *Toi, lumière.* (TPB-11)
Sortie : {(1, « Toi »), (2, « , »), (3, « lumière »), (4, « . »)}

Entrées et sorties du tokenizer

Le *tagger* (ou morphologiseur) est le module qui réalise les prédictions de partie du discours et les caractéristiques morphologiques : les étiquettes *upos* et *xpos*, c'est-à-dire les étiquettes de parties du discours universelles et spécifiques à la langue, ainsi que des informations morphologiques (genre, nombre, mode…) sont prédites pour chaque token.

[18] https://spacy.io/usage/spacy-101.

Entrée : {(1, « Toi »), (2, « , »), (3, « lumière »), (4, « . »)}
Sortie (convertie au format CoNLL-U) :

```
1	Toi	_	PRON	PRON	Number=Sing|Person=2|PronType=Prs	_	_	_	_
2	,	_	PUNCT	PUNCT	_	_	_	_	_
3	lumière	_	NOUN	NOUN	Gender=Fem|Number=Sing	_	_	_	_
4	.	_	PUNCT	PUNCT	_	_	_	_	_
```

Entrées et sorties du morphologiseur

Le *parser* est responsable de la prédiction de l'arbre syntaxique de la phrase. Dans le cas de spaCy, la prédiction du parser détermine par défaut également la division en phrases du texte (→ 4.2.14.2) ; il occupe donc également le rôle de *sentencizer*. Pour chaque token, le modèle prédit l'ID du gouverneur ainsi que l'étiquette de la relation. Par exemple, le modèle prédit dans notre illustration que le token 3 (*lumière*) dépend du token 1 (*Toi*) selon une relation typée *appos* (apposition). Le parser est un modèle statistique entraîné, basé dans notre cas sur une structure de type BERT — celle de CamemBERT. Les étiquettes utilisées par les modèles statistiques dépendent des données sur lesquelles ils ont été entraînés ; dans notre cas, les étiquettes sont donc, comme nous l'avons vu, les étiquettes UD présentes dans le segment Sequoia-Train.

Entrée : {(1, « Toi »), (2, « , »), (3, « lumière »), (4, « . »)}
Sortie (convertie au format CoNLL-U) :

```
1	Toi	_	_	_	_	0	ROOT	_	_
2	,	_	_	_	_	1	punct	_	_
3	lumière	_	_	_	_	1	appos	_	_
4	.	_	_	_	_	1	punct	_	_
```

Entrées et sorties du parser

Le module *ner* (*named-entity recognition*) a pour objet le repérage des entités nommées. D'autres modules standards ou personnalisés peuvent également être ajoutés.

Entrée : *Toi, lumière.*
Sortie (convertie au format CoNLL-U) :

```
1	Toi	toi	PRON	PRON	Number=Sing|Person=2|PronType=Prs	0	ROOT	_	_
2	,	,	PUNCT	PUNCT	_	1	punct	_	_
3	lumière	lumière	NOUN	NOUN	Gender=Fem|Number=Sing	1	appos	_	_
4	.	.	PUNCT	PUNCT	_	1	punct	_	_
```

Entrées et sorties de la chaîne de traitement[19]

2.2.2.1 fr-dep-news-trf

La chaîne de traitement disponible nativement qui nous intéresse particulièrement est la chaîne *fr-dep-news-trf*[20]. Il s'agit de la chaîne de traitement pour le français la plus

[19] Nous avons également considéré la sortie du lemmatiseur, présent par défaut dans les chaînes de traitement, mais n'étant pas illustré dans la figure 4. Il prend la sortie du morphologiseur en entrée et prédit le lemme de chaque token.

[20] https://spacy.io/models/fr#fr_dep_news_trf.

précise dont dispose spaCy. Comme l'indique son nom, il s'agit d'une chaîne de traitement pour le français (*fr*), adoptant la syntaxe dépendancielle (*dep*), développée à l'aide d'un corpus incluant des articles journalistiques (*news*) et soutenue par des transformers (*trf*, il s'agit ici de CamemBERT). Les modèles statistiques de la chaîne ont été entraînés à l'aide du treebank UD-Sequoia (→ 2.1.3.1).

Les composants inclus sont, dans l'ordre (spaCy, s. d.) : *transformer*[21], morphologiseur, parser, *attribute-ruler*[22] et lemmatiseur. Le morphologiseur, l'attribute-ruler et le lemmatiseur ne partageant pas leurs informations avec le parser, qui ne dépend que du module transformer, nous n'évaluons pas la précision de leurs prédictions puisqu'elles ne conditionnent pas la précision de l'analyse syntaxique.

Il est important de noter que, bien que *fr-dep-news-trf* soit basé sur CamemBERT, spaCy a décidé de connecter la dernière couche des encodeurs avec un parser basé sur la transition (*transition-based parser*) plutôt que la tête *biaffine graph-based* sélectionnée à l'origine (Martin *et al.*, 2020 : 7207). Dozat et Manning expliquent la différence entre les approches :

> Transition-based parsers—such as shift-reduce parsers—parse sentences from left to right, maintaining a "buffer" of words that have not yet been parsed and a "stack" of words whose head has not been seen or whose dependents have not all been fully parsed. At each step, transition-based parsers can access and manipulate the stack and buffer and assign arcs from one word to another. One can then train any multi-class machine learning classifier on features extracted from the stack, buffer, and previous arc actions in order to predict the next action. […] At each step, the (feedforward) network assigns a probability to each action the parser can take based on word, tag, and label embeddings from certain words on the stack and buffer. Transition-based parsing processes a sentence sequentially to build up a parse tree one arc at a time. Consequently, these parsers don't use machine learning for directly predicting edges; they use it for predicting the operations of the transition algorithm. Graph-based parsers, by contrast, use machine learning to assign a weight or probability to each possible edge and then construct a maximum spanning tree (MST) from these weighted edges. (Dozat et Manning, 2017 : 1-2)

Cette différence est importante pour comprendre la façon dont la division en phrases est effectuée par le parser : en plus de l'assignation de dépendances, l'algorithme de transition est susceptible de prédire une opération menant à la création

[21] Il s'agit d'un module permettant l'usage par les autres composants des transformers préentraînés de CamemBERT.

[22] Ce module permet la gestion de diverses exceptions prédéfinies : il permet d'ajouter des attributs à des ensembles de tokens correspondant à des configurations choisies.

de plusieurs graphes disjoints (qui ne sont pas connexes) lorsqu'il s'agit de l'opération possédant la plus grande valeur. La propriété de connexité du graphe syntaxique est donc essentielle pour la division en phrases. La division phrastique est donc réalisée au fur et à mesure de l'analyse plutôt que préalablement à celle-ci, lorsque le résultat de l'étape indique que cette division est nécessaire, car plus aucune dépendance ne lie les tokens précédents et les tokens suivants.

Dans le cadre de ce travail, c'est cette chaîne de traitement qui a été sélectionnée, car sa précision à la hauteur des standards industriels, du moins sur des textes journalistiques, est susceptible de permettre les conclusions les plus intéressantes avec notre corpus littéraire. Nous nous attendons cependant à des difficultés dues à la différence de genre textuel.

2.3 Évaluation de parsers : état de l'art et méthodologie

Dans cette section, nous discutons l'évaluation des parsers en TAL. Nous présentons d'abord l'état de l'art de l'évaluation (→ 2.3.1), puis la méthode que nous proposons dans ce travail (→ 2.3.2).

2.3.1 État de l'art de l'évaluation en TAL

Lors de la production d'un outil de traitement automatique des langues, afin de connaître ses performances, en particulier en vue d'une comparaison avec des outils précédemment développés, l'évaluation est primordiale ; le parsing dépendanciel ne déroge pas à la règle. Se sont alors développés différents moyens d'évaluation dont la dichotomie principale, comme le relèvent Resnik et Lin (2010 : 271), est celle qui oppose les évaluations automatiques des évaluations manuelles. Évidemment, la solution qui semble la plus immédiate pour mesurer à quel point un outil est capable de produire les résultats susceptibles d'aider des individus est son utilisation par un ou plusieurs opérateurs humains appliquant un système de notation ; cela apporte cependant différents problèmes comme le relèvent les auteurs (Resnik et Lin, 2010 : 273) : les résultats sont souvent très hétérogènes, car les jugements humains sont sujets à la variation. Dans ce cas, l'évaluation est extrêmement lente et coûteuse. Dans une discipline évoluant très rapidement, il est important de pouvoir tester les outils aisément afin de valider des expériences successives ; c'est donc vers les évaluations automatiques que les chercheurs se sont le plus souvent tournés.

En ce qui concerne les parsers syntaxiques, cette évaluation automatique consiste en, comme le dit Dekang (2003 : 319), une comparaison unité à unité d'un arbre déjà réalisé avec l'arbre prédit par le modèle sur base de la même phrase, d'abord souvent en considérant l'arbre de dépendance sans les étiquettes des relations puis en les prenant en compte (UAS, *unlabeled assignment score* et LAS, *labeled assignment score*). L'exactitude de la prédiction est alors mesurée selon des critères classiques de l'intelligence artificielle : la précision, qui évalue la proportion de résultats corrects (vrais positifs, *tp*) d'une étiquette par rapport au nombre total d'objets désignés par le modèle par cette étiquette (vrais positifs et faux positifs, *tp*+*fp*), donc $p = \frac{tp}{tp+fp}$; le *recall*, ou sensibilité, qui évalue la proportion de résultats corrects (*tp*) d'une étiquette par rapport au total attendu de cette étiquette (vrais positifs et faux négatifs, *tp*+*fn*), donc $r = \frac{tp}{tp+fn}$; le Fscore, moyenne pondérée des scores précédents (Sokolova et Lapalme, 2009 : 430).

Ces scores ne sont évidemment pas limités à la cotation d'arbres entiers, mais sont également attribués pour chaque type de relation : il est alors aisé pour les chercheurs de saisir quelles sont les relations obtenant les meilleurs scores ainsi que les moins bons et agir en conséquence. Cependant, les scores sont dépendants du corpus et des choix réalisés dans celui-ci :

> As it turns out, however, such evaluation procedures are sensitive to the annotation choices in the data on which the parser was trained. Different annotation schemes often make different assumptions with respect to how linguistic content is represented in a treebank (Rambow, 2010). The consequence of such annotation discrepancies is that when we compare parsing results across different experiments, even ones that use the same parser and the same set of sentences, the gap between results in different experiments may not reflect a true gap in performance, but rather a difference in the annotation decisions made in the respective treebanks. (Tsarfaty *et al.*, 2011 : 385)

Une solution serait de sélectionner un seul ensemble permettant de comparer les performances des parsers, mais cela induit nécessairement que les modèles entraînés sur cet ensemble réalisent de meilleurs scores sans pour autant être plus efficaces ou précis lorsqu'ils sont confrontés à des données issues d'une application pratique. Cette évaluation sur un corpus dit *gold standard*, procédé désormais classique pour la publication d'articles concernant des outils de TAL, est durement critiquée par Kovář, Jakubíček et Horák (2016 : 540-541), qui y voient de nombreux inconvénients : la

création d'un corpus *gold* est coûteuse puisqu'elle nécessite le travail d'un spécialiste durant parfois plusieurs mois ; les scores peuvent être biaisés par des incohérences d'annotation entre le corpus d'entraînement et le corpus d'évaluation, et cela entraîne un processus de surajustement (*overfitting*) des modèles à ces standards qui ne sont pas toujours représentatifs de la réalité de l'usage pratique.

> However, the gold standard enforces that all tools need to solve all the problems covered by the gold standard and in the same way as the gold standard prescribes (e.g. with the same granularity), otherwise they will lack a sound evaluation according to the state-of-the-art methodology. This way, the NLP tools are designed according to the gold standard "shape". (Kovář, Jakubíček et Horák, 2016 : 541)

Nivre et Fang (2017 : 86) ajoutent que cette évaluation quantitative automatique du parsing syntaxique ne met pas sur un pied d'égalité toutes les langues et favorise les langues analytiques, ce qui risque de dévaluer des architectures audacieuses, mais développées pour des langues plus synthétiques.

2.3.2 L'évaluation proposée dans ce travail

Plutôt qu'une évaluation purement automatique et basée sur des gold standards, ce travail propose donc d'analyser les résultats d'une analyse syntaxique d'un texte littéraire en replaçant les considérations linguistiques au centre de la réflexion grâce à un examen approfondi. En procédant comme ceci, l'objectif n'est pas de comparer les performances de différents parsers puisque nous ne nous centrons que sur le modèle *fr-dep-news-trf* de spaCy, mais il est plutôt de considérer, en vue d'une éventuelle application littéraire de l'analyse, les forces et faiblesses linguistiques de l'analyse. Les conclusions peuvent donc dépasser le parser et permettre de se prononcer sur des particularités syntaxiques du corpus de test, mais aussi du corpus d'entraînement.

La méthodologie de notre évaluation se divise en deux phases : **1.** la préparation des données avant la soumission au parser, en ce compris le choix du corpus et du parser ; **2.** l'annotation et l'analyse des résultats.

2.3.2.1 Phase préparatoire

Il s'agit de s'interroger sur le choix du corpus et le mode de présentation des données au parser. Le choix du parser a déjà été expliqué auparavant.

La **sélection du corpus** est motivée par deux paramètres : l'intérêt de l'étude à l'analyse d'un texte d'un genre différent de celui de la communication grand public, ce

qui a conduit au choix d'un corpus issu de la littérature, à vocation artistique, et des préférences personnelles.

Le choix d'un texte littéraire est particulièrement intéressant, car il s'inscrit dans la réflexion sur « ce qui fait la littérarité d'un texte » en plus de proposer, par son aspect artistique, des constructions syntaxiques originales et peu courantes dans la langue commune de tous les jours. La recherche de complexité syntaxique nous a mené à la poésie en prose, en particulier *Flaques de verre* de Pierre Reverdy (1929), recueil qui touche notre sensibilité. Il s'agit d'un recueil présentant des surprises sémantiques régulièrement traduites sur le plan syntaxique (à l'image de PT-4, par exemple) sans pour autant entrer en rupture totale avec la norme du français écrit, ce qui aurait appauvri nos résultats.

> (PT-4) Tout entière, la région tourne au fil du cadran — les élans des rayons dorés sous la paupière doublés par le bruit sourd de la vie des rivières et des pentes gardées par des plis remuants — jusqu'aux franges du ciel où fument les prières, dans la campagne grise et les cris du couchant.

En ce qui concerne la taille de l'échantillon, nous avons décidé de sélectionner un tiers des pièces en sélectionnant un poème sur trois à partir du premier. Nous obtenons alors un corpus de vingt-cinq poèmes sur les septante-quatre que compte le recueil, pour un total de 4353 tokens répartis en 219 phrases, ce qui nous semble à la fois raisonnable — l'objectif n'étant pas une analyse uniquement quantitative, mais aussi une approche qualitative de matériaux singuliers et concrets — et représentatif autant en ce qui concerne la proportion que la répartition. Ces poèmes ont été transcrits en un format numérique en utilisant des méthodes de reconnaissance optique[23] et corrigés à la main. Le texte est issu de l'édition Flammarion (GF) de 1984.

Afin de limiter au maximum les erreurs de division en phrases, nous avons décidé de soumettre les phrases, délimitées à la main grâce à la ponctuation, indépendamment les unes après les autres. Bien que nous perdions la possibilité de nous interroger sur la notion de phrase chez Reverdy, qui intègre régulièrement des signes forts de ponctuation, comme le point, entre des éléments qui semblent être sémantiquement (voire syntaxiquement) liés, le nombre de relations ignorées à cause d'erreurs de la division phrastique risquait d'être trop important et un effet cumulatif des erreurs à la

[23] https://github.com/tesseract-ocr/tesseract.

suite d'une première division erronée était à attendre. Un script écrit dans le langage de programmation Python soumet donc les phrases à la chaîne de traitement *fr-dep-news-trf* et compile les sorties en un unique fichier au format CoNLL-U qui reprend les résultats de l'analyse. Bien que les informations de parties du discours soient disponibles dans ce fichier de sortie, nous les ignorons, car elles dépendent du morphologiseur qui n'est pas lié au parser.

2.3.2.2 Annotation et évaluation des résultats

C'est ce fichier qui constitue la base du reste du travail. Grâce à un logiciel développé par nous pour l'occasion, nous annotons chaque erreur repérée dans les données de sortie en la corrigeant et en lui assignant un code permettant une première typologie très sommaire, illustrée par la figure 5 : HEAD (flèches en magenta) signifie que la prédiction de l'étiquette de la relation est correcte, mais que la prédiction du gouverneur est incorrecte ; DEPREL (flèches en cyan) signifie que le gouverneur est correct, mais que l'étiquette est incorrecte ; BOTH (flèches en rouge) signifie qu'à la fois l'étiquette et le gouverneur sont incorrects.

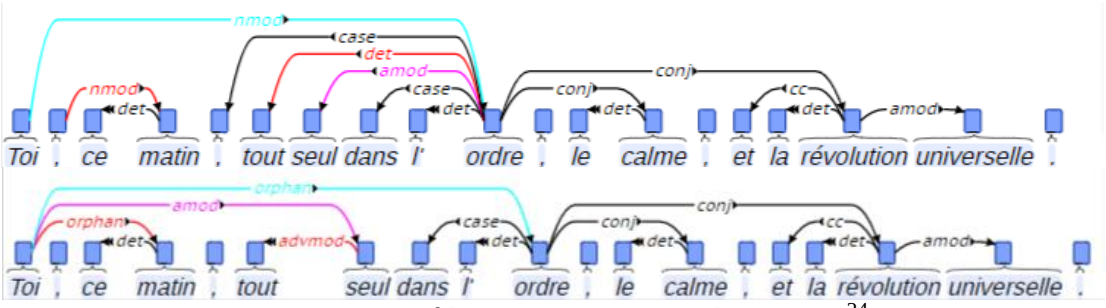


Fig. 5 — Prédiction et correction de TPB-15[24]

Le code SEG indique une erreur de segmentation interne de la phrase (fig. 6, en vert ; la prédiction est constituée de deux graphes disjoints). Nous pourrions considérer que SEG est un sous-type de BOTH : puisque le segment n'est pas considéré comme dépendant d'un autre, et que cette relation n'est donc pas étiquetée, le gouverneur et l'étiquette sont nécessairement incorrects. Cependant, puisque cette erreur se produit

[24] Ces visualisations d'arbres syntaxiques sont réalisées à l'aide d'Annodoc (https://github.com/spyysalo/annodoc). Chaque rectangle bleu correspond à un token, dont la forme est présentée juste en dessous. Les flèches indiquent les relations de dépendance : une flèche partant d'un token indique un dépendant de celui-ci. Elles sont accompagnées des étiquettes de relation. Par exemple, la flèche en magenta du graphe prédit (image supérieure) indique que *seul* dépend d'*ordre* selon une relation d'épithète (*amod*). Toutes les relations ne sont pas systématiquement présentes dans la visualisation lorsque nous voulons en mettre certaines en évidence. Il s'agit par exemple du cas de la ponctuation dans cette phrase.

régulièrement en des lieux remarquables, par exemple dans des situations de parataxe, il nous semblait important de les noter. Ces erreurs sont abordées en détail dans la typologie (→ 4.2.14.2).

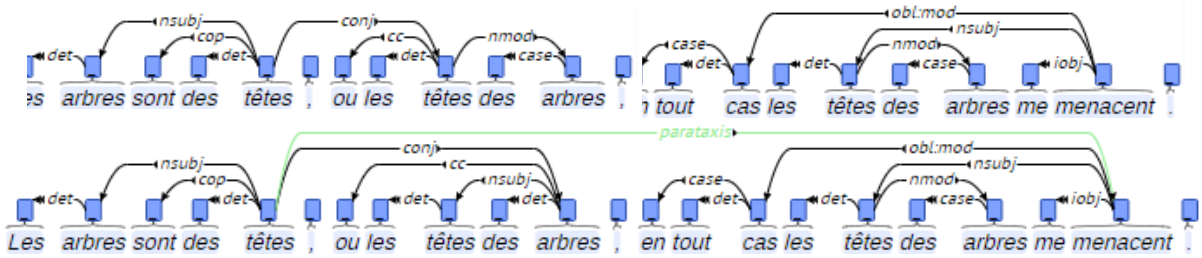


Fig. 6 — Prédiction et correction de C-2

Les erreurs les plus communes sont BOTH (173 occurrences), suivies de DEPREL (155), HEAD (72) et SEG (23).

En ce qui concerne la ponctuation, nous avons décidé de suivre une méthodologie qui nous a été conseillée : puisque la ponctuation est toujours une feuille (un élément terminal) de l'arbre syntaxique, qu'elle ne joue pas de rôle sur le plan syntaxique (si ce n'est celui de donner des indices au lecteur) et que les seules règles qui régissent son annotation dans l'environnement UD sont des recommandations permettant de conserver la projectivité[25] des arbres[26], il nous semble opportun de l'ignorer. En ce sens, nous avons donc fait dépendre automatiquement toute relation d'un token représentant un signe de ponctuation de la racine de l'arbre syntaxique. Nous ignorons de même toute erreur liée à la ponctuation, à l'exception de deux d'entre elles qui sont particulièrement rares et étonnantes : l'une montre un signe de ponctuation en situation de nœud, c'est-à-dire duquel dépend un autre élément ; l'autre montre un signe de ponctuation dont l'étiquette de relation syntaxique n'est pas *punct*. Nous y revenons dans la section 4.2.13.1. Nous avons donc également résolu manuellement les erreurs de segmentation excluant uniquement un signe de ponctuation (voir BT-7), sans les annoter *SEG*.

> (BT-7) Tous les départs et les voyages commencés interrompus et le ciel dépouillé des plus belles étoiles | .

[25] Propriété d'un arbre syntaxique impliquant qu'aucune dépendance ne croise une autre lorsque cet arbre est représenté sous forme linéaire. Nous revenons sur cette notion dans la typologie (→ 4.2.14.2).

[26] https://universaldependencies.org/u/dep/punct.html.

Lorsque les données ont été corrigées et annotées, nous effectuons deux types d'évaluation : une évaluation automatique à l'échelle du corpus (*quantitative,* → 3) et une évaluation des erreurs individuellement (*qualitative,* → 4).

a) Évaluation automatique

Cette évaluation automatique est typique de la méthodologie de l'évaluation en TAL. Elle permet **1.** de donner une idée de l'efficacité du modèle à travers des indicateurs comme le score UAS, le score LAS ; **2.** de considérer individuellement les prédictions des étiquettes grâce au calcul de la précision, de la sensibilité et du F-score de chacune de celles-ci présentes dans le corpus. Cette section nous permet d'illustrer les méthodes automatiques et de mettre en évidence le comportement du modèle à l'égard de certaines étiquettes auxquelles nous devons être attentifs dans le reste du travail.

b) Évaluation individuelle des erreurs

Il s'agit ici de réaliser une typologie linguistique des erreurs. Nous observons alors quelles sont les caractéristiques des différentes catégories d'erreurs, les causes linguistiques éventuelles, leur contexte d'apparition et le comportement du modèle dans la prédiction. Nous divisons d'abord les erreurs selon des paramètres transversaux (erreurs incohérentes, cohérentes et ambiguës) avant de les classer selon une typologie plus précise basée sur les fonctions et structures syntaxiques. Dans chaque section, nous présentons les caractéristiques théoriques de chaque fonction envisagée.

Il s'agit de l'originalité de notre étude : grâce à cette typologie, nous espérons pouvoir expliquer les difficultés rencontrées par le parser et proposer des pistes d'amélioration, soit de l'entraînement, soit du corpus, soit des règles d'annotation UD.

3 Évaluation automatique des résultats du parsing

Cette section présente les résultats d'une évaluation automatique du parser *fr-dep-news-trf* de spaCy. Celle-ci nous permet de connaître les scores UAS[27] (*unlabeled*

[27] Pour rappel, les scores UAS (*unlabeled attachment score*) et LAS (*labeled attachment score*), qui sont très fréquents dans les évaluations automatiques, correspondent à la proportion de tokens dont la prédiction est correcte. Le score UAS, qui ne prend pas en compte les étiquettes (*unlabeled*), considère correct tout token dont le gouverneur est correct. Il ne s'intéresse qu'au squelette syntaxique. Le score LAS, lui, prend en compte les étiquettes (*labeled*) : pour que la prédiction d'un token soit correcte, il faut que la prédiction du gouverneur **et** de l'étiquette de relation soit correcte. Le score LAS est donc toujours inférieur ou égal au score UAS.

attachment score) et LAS (*labeled attachment score*), ainsi que, pour chaque étiquette, la précision (nombre de vrais positifs par rapport au nombre d'occurrences de l'étiquette dans la prédiction), la sensibilité (*recall* ; nombre de vrais positifs par rapport au nombre d'occurrences de l'étiquette dans le corpus de référence) et le F-score (moyenne pondérée des deux précédentes valeurs). Nous avons décidé ici de comparer l'efficacité du parser sur son corpus d'entraînement (Sequoia-Train), le segment de test de Sequoia (Sequoia-Test) et notre corpus littéraire (Reverdy).

La comparaison avec Sequoia-Test permet d'évaluer un possible surajustement aux données du corpus d'entraînement en utilisant des données similaires, du moins du point de vue du genre textuel, à celles que le parser a déjà rencontrées[28]. En effet, si le parser est beaucoup moins efficace sur un corpus de test très similaire à son corpus d'entraînement, nous pouvons mettre en doute sa capacité de généralisation et considérer que le modèle est surajusté, c'est-à-dire que le modèle, à cause d'un entraînement excessif, a perdu en généralisation en s'ajustant trop précisément aux données avec lesquelles il a été entraîné.

Nous présentons tout d'abord les scores LAS et UAS du modèle pour nos corpus (→ 3.1) avant de nous intéresser aux étiquettes en particulier (→ 3.2).

3.1 Efficacité générale du modèle : (*un*)*labeled attachment score*

Le tableau ci-dessous présente l'efficacité sur l'ensemble du corpus grâce aux scores UAS et LAS.

	Sequoia-Train	**Sequoia-Test**	**Reverdy**
UAS	94,1 %	89,5 %	92,4 %
LAS	93,3 %	87,2 %	87,8 %

Tab. 1 — Efficacité du parser sur les corpus

Les résultats sur le corpus d'entraînement sont particulièrement bons : 94,1 % des têtes prédites sont correctes et 93,3 % des tokens sont parfaitement prédits. Cependant, malgré la similarité que nous avons constatée entre Sequoia-Train et Sequoia-Test, la différence d'efficacité est de 4,6 points pour le score UAS et plus de 6 points pour le score LAS. Il semble donc il y avoir ici un cas important de surajustement. Or,

[28] Nous pouvons confirmer que Sequoia-Train et Sequoia-Test sont très similaires, du moins en étudiant la longueur des phrases, la longueur moyenne des dépendances (→ 4.1.2.2), le poids combiné du flux (→ 4.1.2.2) et la fréquence des étiquettes. Puisque ce processus s'éloigne du cadre de notre travail, nous avons décidé de placer ces données en annexe (A.5). Ceci nous permet également de constater qu'il existe un écart selon chacun des indicateurs entre Sequoia et Reverdy.

l'efficacité réelle du parser doit toujours être considérée par rapport à sa capacité de généralisation à de nouvelles données.

Ensuite, nous pouvons observer que l'efficacité du parser est en baisse sur le corpus littéraire par rapport au corpus d'entraînement, mais tout en restant supérieure à l'efficacité sur Sequoia-Test, ce qui semble contre-intuitif : comment un modèle pourrait-il mieux s'adapter à des données présentant des variations liées à une différence de genre textuel qu'à des données très similaires au corpus d'entraînement ? La réponse se trouve probablement dans un biais que nous n'avions pas envisagé avant ce résultat : là où l'évaluation sur Sequoia-Test est réalisée sur un corpus ayant été converti depuis une version en constituants, celle effectuée sur Reverdy est réalisée sur un corpus issu de la correction d'une prédiction du parser évalué. De cette façon, les résultats risquent d'être surévalués : pour tout point flou du modèle UD ou toute ambiguïté syntaxique que nous ne pouvions pas trancher, nous avons pu considérer l'option sélectionnée par le parser comme acceptable alors que nous n'aurions pas forcément répondu de la même façon en annotant à la main. Or, puisque ce score se base sur des paramètres binaires (*vrai* ou *faux* par rapport à la référence) plutôt qu'une échelle de similarité (*plus* ou *moins similaire* par rapport à la référence), des différences minimes qui relèveraient du domaine de l'agrément interannotateur (Arstein, 2017 ; désormais *IAA*) obtiennent le même résultat (*faux*) que des erreurs importantes. La marge d'erreur est alors bien plus importante lors de l'évaluation grâce à Sequoia-Test que lors de celle effectuée sur la correction de la première prédiction de Reverdy qui voit une différence liée à l'IAA qui tend vers 0. Ce biais nous prévient donc de formuler toute conclusion concernant l'efficacité du parser et nous nous contentons de nous intéresser à son comportement face à des difficultés linguistiques.

3.2 Performance de la prédiction individuelle des étiquettes

Le graphique de la figure 7 montre visuellement la variation du F-score (moyenne entre la précision et la sensibilité) des étiquettes dans les trois corpus évalués automatiquement.

Nous pouvons constater quelques étiquettes dont le F-score est égal à 0 dans les trois corpus : *advcl:cleft* (constructions emphatiques de type *c'est… qui*), *aux:caus* (constructions causatives), *csubj* (proposition sujet), *csubj:pass* (proposition sujet passive), *discourse* (marque discursive), *dislocated* (reprise syntaxique, répétition d'un

élément, mais à une position inhabituelle), *goeswith*, *iobj:agent* (pronom complément d'objet indirect correspondant au complément d'agent), *nsubj:caus* (sujet d'une construction causative), *obj*:*agent* (complément d'objet correspondant au complément d'agent) et *orphan*. Il s'agit de relations minoritaires, assez rares, qui posent des problèmes dans tous les corpus étudiés. La plupart du temps, il s'agit de difficultés liées à la faible fréquence de ces étiquettes dans le corpus d'entraînement. Certaines valeurs sont également expliquées simplement par le fait qu'il n'y a aucune occurrence de cette étiquette dans le corpus et donc son F-score vaut automatiquement 0. À part certaines relations comme *orphan* (cas particuliers d'ellipse ; → 4.2.3) et *goeswith* (mots uniques divisés en plusieurs tokens ; → 4.2.2) qui apparaissent respectivement 46 et 12 fois, la plupart des relations n'apparaissent qu'une fois au plus dans notre corpus et il s'agit donc de cas isolés que nous détaillons plus tard. Ces étiquettes ne sont pas significatives pour définir le comportement du parser, mais montrent quand même certaines faiblesses liées aux relations extrêmement rares.

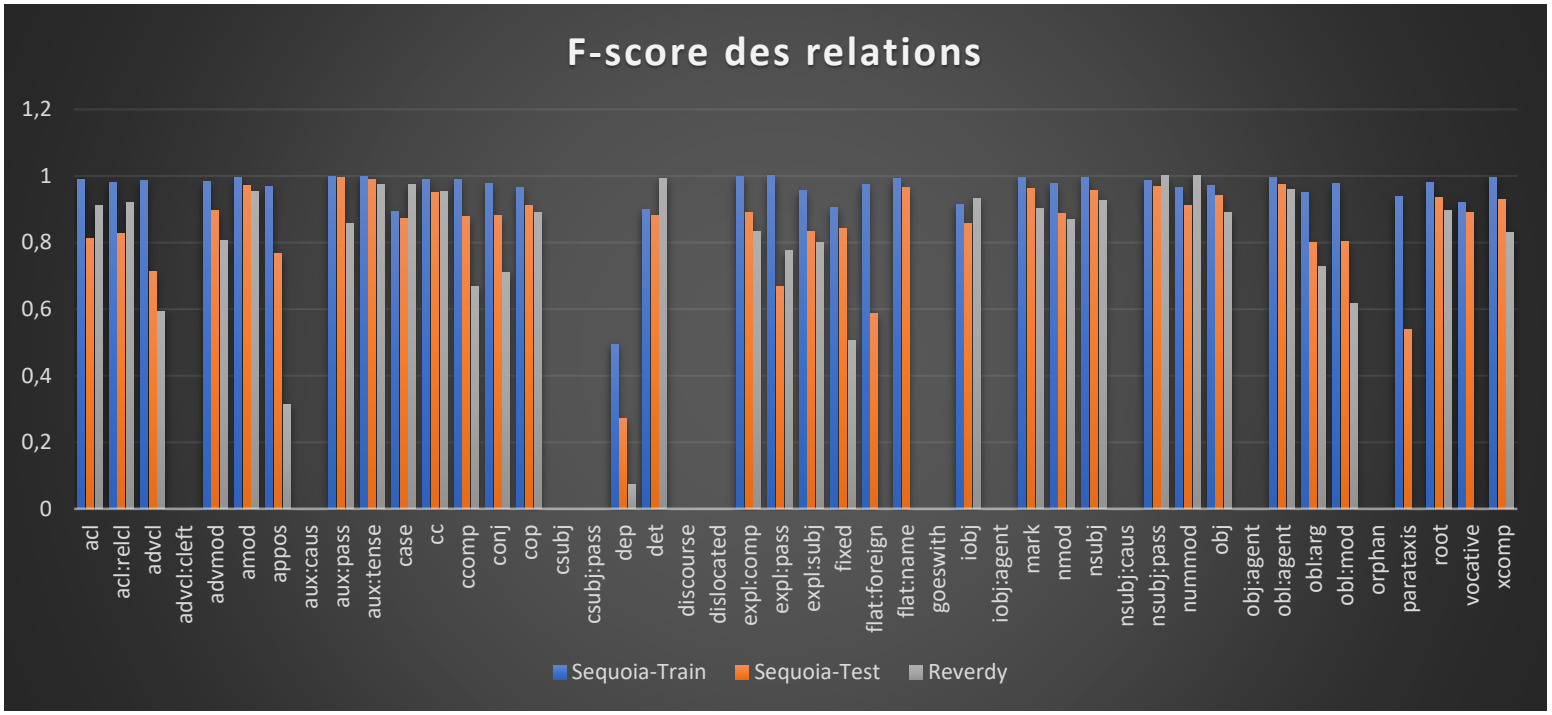


Fig. 7 — F-score des étiquettes syntaxiques dans l'évaluation automatique des corpus

Les étiquettes *appos* (apposition), *dep* (dépendance non spécifiée), *fixed* (utilisée pour les figements) et *parataxis* (parataxe) ont un F-score très faible et en forte baisse par rapport aux corpus Sequoia. Il est donc important de leur réserver une place importante dans la typologie des erreurs et de nous intéresser aux raisons (linguistiques

ou techniques) derrière ces difficultés. L'efficacité des relations *flat:foreign* (séquence de mots étrangers) et *flat:name* (séquence de noms propres), qui ont un F-score de 0 pour l'évaluation avec Reverdy malgré des performances élevées dans Sequoia, est expliquée par le fait qu'aucune occurrence n'est rencontrée.

En multipliant la marge d'erreur (calculée à partir du F-score) par le nombre d'occurrences, nous obtenons la figure 8, permettant de considérer le nombre absolu d'erreurs attendues durant la correction de la prédiction sur le corpus littéraire. Les relations qui se démarquent ne sont pas toujours les mêmes que celles envisagées supra : *conj* (conjoint d'une coordination) et *obl:mod* (complément adverbial non essentiel) produisent le plus d'erreurs, suivies d'*orphan*, *nmod* (complément déterminatif), *advmod* (modifieur adverbial), *root* (racine de la phrase), *fixed* et *appos*. Puisqu'il s'agit d'un concept complexe du point de vue de la théorie de la grammaire dépendancielle, nous consacrons, dans la suite de ce travail, un point complet à la coordination, dont *conj* (→ 4.2.1). Il est très important également de nous intéresser aux cas d'ellipse au vu du nombre d'erreurs attendues liées à *orphan*.

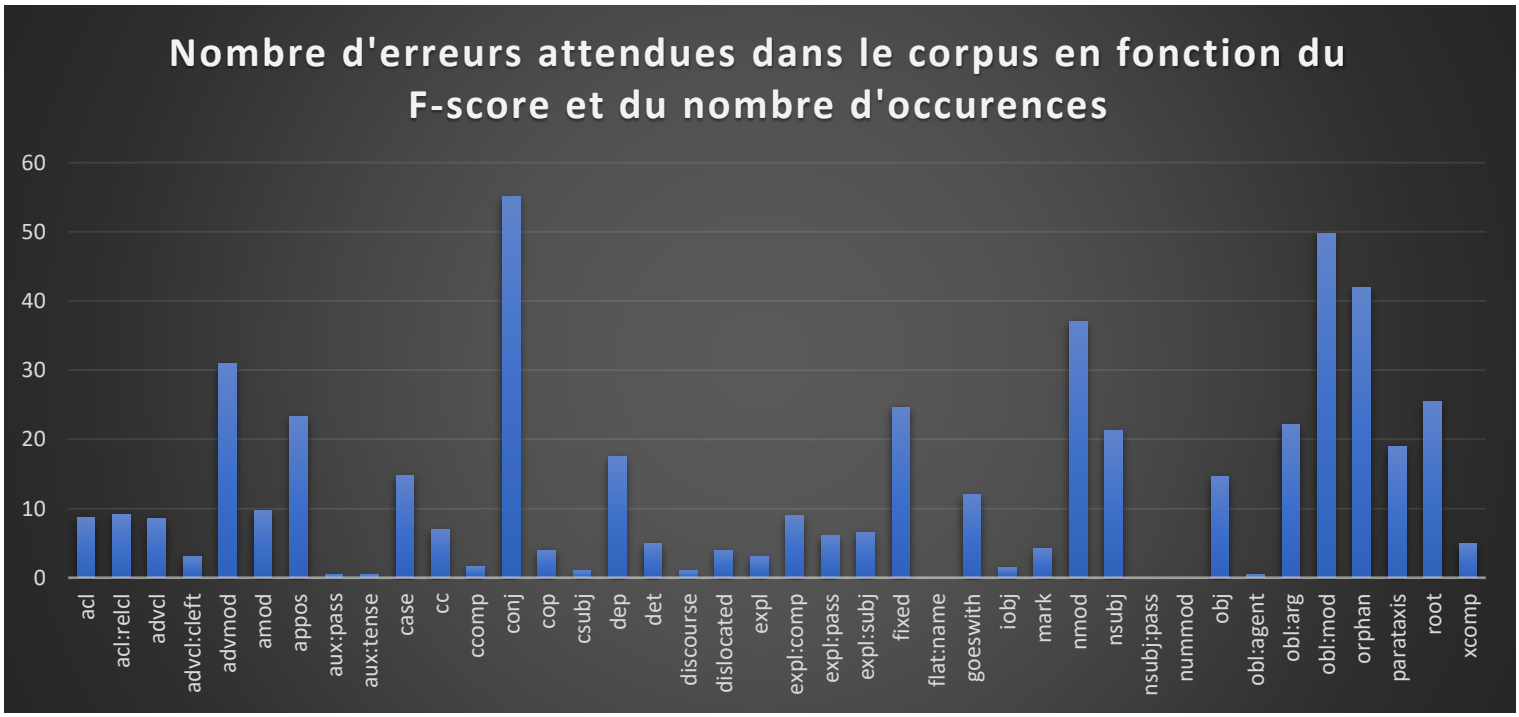


Fig. 8 — Nombre absolu d'erreurs attendues pour chaque étiquette

La figure 9 présente la différence entre la précision et la sensibilité de la prédiction pour chaque relation. La précision n'étant pas impactée par les faux négatifs et la sensibilité par les faux positifs, nous pouvons dégager des tendances d'annotation pour

les étiquettes : lorsque la précision est plus élevée que la sensibilité (valeurs positives sur le graphique), le modèle a une tendance aux faux négatifs. Dans le cas contraire (valeurs négatives), le modèle a un comportement plutôt naïf (faux positifs).

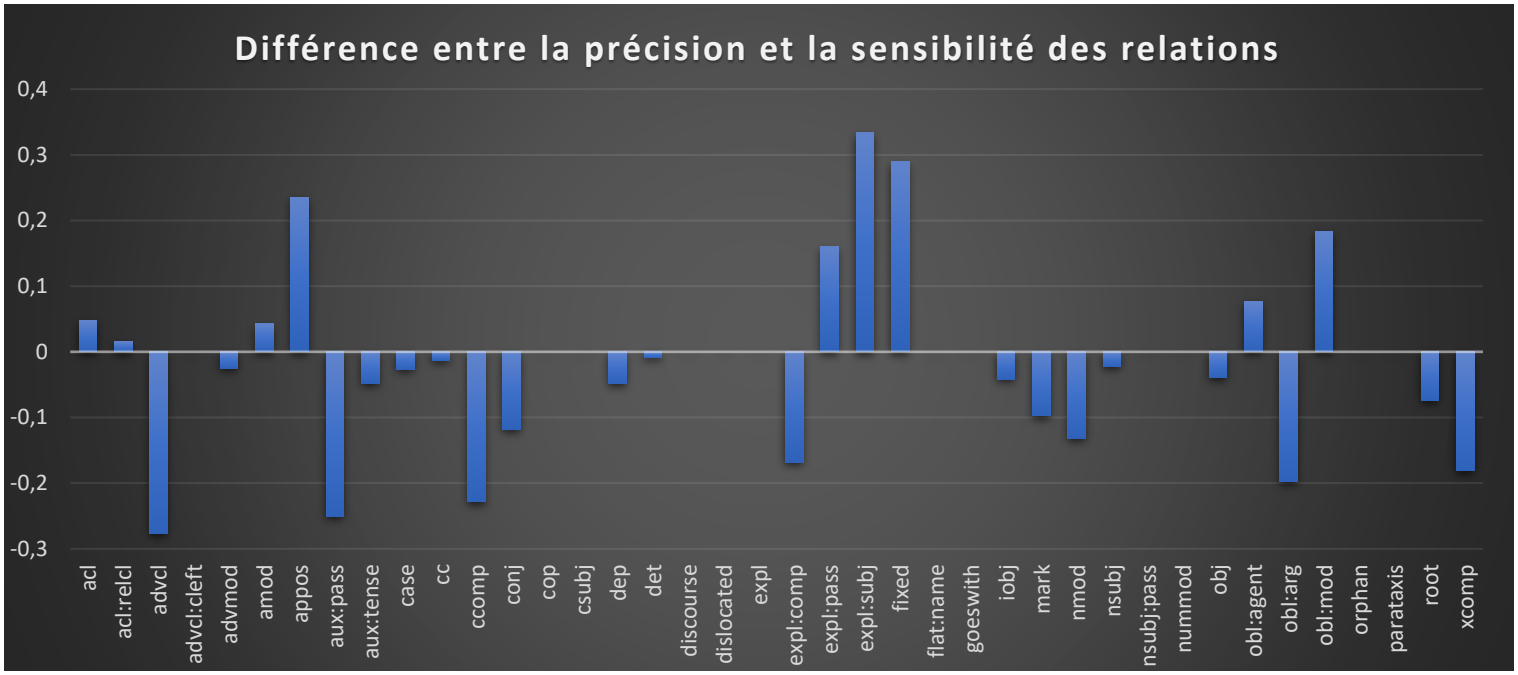


Fig. 9 — Différence entre la précision et la sensibilité dans l'évaluation des relations de Reverdy

Nous pouvons donc considérer qu'une grande partie des erreurs liées aux étiquettes *appos*, *expl:pass*, *expl:subj*, *fixed* et *obl:mod* sont des cas dans lesquels une autre étiquette, erronée, a été sélectionnée (faux négatifs) au lieu de l'étiquette correcte, induisant une sous-représentation de ces étiquettes. Dans le cas de *advcl*, *aux:pass*, *ccomp*, *obl:arg*, *xcomp* et *nmod*, le modèle a plutôt tendance à sélectionner ces étiquettes dans des situations inadéquates, ce qui conduit à une surreprésentation de ces étiquettes. Ces résultats sont pris en compte dans l'analyse accompagnant la typologie linguistique (→ 4.2).

Le recensement des paires d'étiquettes par fréquence est une étape que nous avons décidé d'ajouter à cette évaluation automatique, car la présence de récurrences dans la substitution des étiquettes peut être un indicateur particulièrement utile pour la suite de l'analyse. Il s'agit de relever, pour chaque prédiction d'étiquette erronée (erreurs DEPREL et BOTH), quelles sont l'étiquette prédite et l'étiquette correcte, qui ont donc été confondues. Nous obtenons ainsi une liste de paires d'étiquettes qui sont

régulièrement confondues[29] (tab. 2). Par exemple, dans *Pierre mange une pomme*, si *Pierre* a été prédit comme dépendant de *mange* avec l'étiquette de complément d'objet direct *obj* (« *Pierre* <obj *mange*[30] »), alors qu'il s'agit en réalité du sujet (*nsubj* ; « *Pierre* <nsubj *mange* »), nous comptons une occurrence de la paire (*nsubj*, *obj*).

Étiquette A	Étiquette B	Nombre d'occurrences de la paire	A est corrigé en B (A -> B)	B est corrigé en A (B -> A)
obl:arg	obl:mod	23	19	4
nmod	orphan	22	22	0
appos	nmod	12	0	12
dep	obl:arg	11	4	7
nmod	obl:arg	11	8	3
nmod	obl:mod	11	9	2
case	fixed	8	8	0
conj	obl:mod	7	6	1
advmod	fixed	6	6	0
expl:subj	nsubj	6	0	6

Tab. 2 — Paires des étiquettes prédiction-correction les plus fréquentes des étiquettes

Il est important de noter que nous ne différencions pas, dans le compte du nombre d'occurrences, les situations dans lesquelles *nsubj* est prédit et *obj* est correct des situations dans lesquelles *obj* est prédit et *nsubj* est correct. En effet, l'étiquette A est simplement celle apparaissant en premier dans l'ordre alphabétique. Dans un second temps, nous comptons le nombre de situations dans lesquelles l'étiquette A est corrigée par l'étiquette B (**A -> B** ; A, prédite, est incorrecte et B, corrigée, est correcte) et l'étiquette B est corrigée par l'étiquette A (**B -> A** ; B, prédite, est incorrecte et A, corrigée, est correcte). Cela nous permet de repérer quelles sont les paires fréquemment confondues, et donc d'estimer une certaine proximité dans l'annotation syntaxique perturbant le modèle, mais aussi de mettre en évidence le comportement du modèle dans certaines paires : quelle étiquette est-elle le plus souvent en situation de faux positif, ou de faux négatif au sein d'une paire ?

[29] Nous incluons les dix paires les plus fréquentes, ce qui correspond à 35 % des confusions. La liste exhaustive se trouve en annexe (M.2).

[30] Nous utilisons la notation de Kahane, Yan et Botalla (2017), qui conserve la linéarité des tokens tout en indiquant le sens de la dépendance. « *a* <rel *b* » indique que *a* est dépendant de *b* selon la relation étiquetée *rel* ; « *a* rel> *b* » indique que *b* est dépendant de *a* selon cette même relation.

Nous pouvons remarquer que les paires fréquentes incluent généralement des étiquettes partageant des caractéristiques syntaxiques : *obl:mod* et *obl:arg* ne se différencient qu'au niveau de l'essentialité du complément adverbial ; *nmod* et *obl:arg* sont toutes deux des relations de syntagmes prépositionnels… La paire (*appos*, *nmod*) semble contre-intuitive : bien qu'il s'agisse dans les deux cas de dépendants d'un substantif (*mon chat, **ce félin*** et *le chat **du voisin***), l'apposition est constituée par un syntagme nominal alors que le complément déterminatif est un syntagme prépositionnel. Nous considérons ce cas, qui est pour nous une conséquence du choix des têtes lexicales dans UD, en particulier dans la section des appositions (→ 4.2.1.3).

La paire (*nmod*, *orphan*), très fréquente, est beaucoup moins évidente à comprendre sans observer les données ; la suite de ce travail intègre donc des considérations sur la confusion entre le complément déterminatif et les dépendants d'un élément passé sous ellipse (→ 4.2.3).

Grâce à ceci, nous pouvons également mettre en évidence les faux positifs : *nmod* est très régulièrement faux positif, qu'il s'agisse des paires (*nmod*, *orphan*), (*appos*, *nmod*), (*nmod*, *obl:arg*) et (*nmod*, *obl:mod*) ; d'autre part, *obl:arg* est en majorité un faux positif dans la paire (*obl:arg*, *obl:mod*).

3.3 Conclusion de l'évaluation automatique

Dans cette section, nous avons pu isoler une série d'étiquettes problématiques pour le modèle, qu'il s'agisse de faux négatifs ou de faux positifs. Nous avons également pu mettre en évidence certaines paires d'étiquettes fréquemment confondues, ce qui peut indiquer des difficultés liées à la similarité, dans UD, de certaines relations syntaxiques. Les étiquettes produisant le plus d'erreurs à l'échelle du corpus semblent être les étiquettes *conj* (coordination), *obl:mod* (complément adverbial non essentiel), *orphan* (ellipse) et *nmod* (complément déterminatif). Cette dernière est souvent prédite en confusion avec *appos* (apposition), mais aussi *orphan*, *obl:arg* et *obl:mod*. Enfin, la proximité d'*obl:mod* et *obl:arg*, relations distinguées uniquement du point de vue de l'essentialité du complément adverbial, semble problématique pour le modèle puisqu'il s'agit de la paire la plus fréquemment confondue.

Passons désormais à une évaluation selon des paramètres linguistiques plutôt que statistiques.

4 Évaluation linguistique et manuelle de la prédiction

Cette section constitue le cœur de notre travail et est le lieu de son originalité. Il s'agit ici de nous écarter des évaluations automatiques qui semblent bien plus appartenir au domaine des statistiques que de celui de la linguistique, duquel est pourtant issu le matériau sur lequel est réalisée la prédiction.

Dans un premier temps, nous nous intéressons à des paramètres transversaux qui peuvent jouer un rôle dans toute prédiction erronée (→ 4.1). Nous décrivons ainsi une première typologie, grossière, basée sur la cohérence de la prédiction. Nous nous intéressons ensuite au contexte d'apparition de ces erreurs selon des considérations de longueur et de flux de dépendance. Nous pourrions qualifier cette première sous-section de *linguistique quantitative* : il s'agit là de nous interroger sur des paramètres influençant les erreurs dans leur totalité.

Dans un second temps, nous proposons une typologie des erreurs basée sur des critères linguistiques comme des fonctions ou des phénomènes syntaxiques (→ 4.2). À l'aide d'une évaluation manuelle de chaque erreur, nous tentons de dégager une série de facteurs structurants expliquant les difficultés rencontrées par le parser, qu'elles soient liées aux structures rencontrées, à des conventions de UD ou même à des phénomènes linguistiques, comme l'ambiguïté, auxquels un parser est inévitablement soumis. Pour chaque section, nous décrivons le comportement du parser face aux situations concernées et formulons des hypothèses d'explication de ce comportement afin de cerner les faiblesses de la prédiction, et même de UD. Évidemment, au vu de la complexité de certaines des erreurs que nous avons rencontrées, les sections que nous dégageons ne sont pas des ensembles fermés : il est extrêmement courant qu'une erreur puisse relever d'une multitude de phénomènes. Chaque section est nommée selon un phénomène (par exemple l'ellipse, l'homonymie ou les relations d'équivalence) autour duquel semblent graviter une certaine quantité d'erreurs de la prédiction. Ainsi, la section *Ellipse* (→ 4.2.3) reprend toutes les erreurs qui relèvent à nos yeux de l'omission d'un élément.

4.1 Paramètres transversaux

Dans cette section, l'objectif est de nous intéresser aux erreurs selon un point de vue quantitatif, mais fondé sur des considérations linguistiques comme celles de cohérence

de la prédiction (→ 4.1.1) ou de complexité linguistique, celle-ci étant notamment liée au flux de dépendance selon Kahane, Yan et Botalla (2017 ; → 4.1.2).

4.1.1 (In)cohérence de la prédiction : une première typologie des erreurs

La première division que nous proposons est centrée autour de la notion de cohérence : en quelle mesure une erreur est-elle justifiée du point de vue syntaxique ? Pour nous, en effet, une prédiction erronée peut avoir différents niveaux de *motivation syntaxique*, c'est-à-dire qu'elle peut être plus ou moins justifiée (***cohérente***) par rapport aux données présentées au parser.

Afin de tenir compte des difficultés techniques de la prise en compte du contexte dans son ensemble, nous considérons distinctement la cohérence aux échelles microscopique et macroscopique.

Par microscopique, nous entendons la relation entre le gouverneur et le dépendant : est cohérente sur le plan microscopique une prédiction qui est correcte lorsque nous isolons le gouverneur et le dépendant du reste de la phrase. Par exemple, la prédiction de la dépendance « *mange* nmod> *pomme* » dans *je mange une pomme* est incohérente à l'échelle microscopique puisque l'étiquette *nmod* ne peut être utilisée pour un dépendant verbal. Dans *Il m'annonce à son maître*, les dépendances « *m'* <obj *annonce* » et « *m*' <iobj *annonce* » sont cohérentes toutes les deux puisque le pronom *me* peut être autant complément d'objet direct que complément d'objet indirect et le verbe *annoncer* est transitif direct et indirect, c'est-à-dire qu'il accepte trois positions actancielles (sujet, complément d'objet direct, complément d'objet indirect). Nous ne considérons pas ici le fait que, en considérant la phrase **dans son entier,** le pronom ne peut être complément d'objet indirect, car cet actant est déjà réalisé par le syntagme prépositionnel *à son maître*. Notons qu'à l'échelle microscopique nous considérons la présence ou l'absence des mots fonctionnels : « *Philippe* nmod> *roi* » dans *Philippe, roi des Belges* est incohérent puisque l'étiquette *nmod* en français ne peut s'utiliser que pour des syntagmes prépositionnels.

Au contraire, par macroscopique, nous entendons l'arbre syntaxique dans sa globalité : la prédiction est cohérente à l'échelle macroscopique si elle est correcte dans le contexte entier de la phrase. Ainsi, dans l'exemple précédent (*Il m'annonce à son maître*), « *m'* <obj *annonce* » est cohérente sur le plan macroscopique (puisqu'une place est disponible pour le complément d'objet direct, que nous remplissons par *me*) alors

que « *m'* <iobj *annonce* », cohérent à l'échelle microscopique, est incohérente à l'échelle macroscopique puisque le complément d'objet indirect est déjà réalisé par un autre syntagme.

À nos yeux, une primauté peut être accordée au contexte microscopique puisqu'une erreur cohérente à l'échelle macroscopique, mais incohérente à l'échelle microscopique nous semble être uniquement le fruit du hasard ; nous dégageons alors trois positions :

- les erreurs ambiguës, prédictions cohérentes dans un contexte microscopique, mais aussi macroscopique, pour lesquelles *l'erreur*[31] se situe au-delà de la syntaxe de la phrase (sens, parallélisme textuel…) ;
- les erreurs cohérentes, prédictions cohérentes dans un contexte microscopique, mais incohérentes dans un contexte macroscopique ;
- les erreurs incohérentes, prédictions incohérentes dans un contexte microscopique (et, *a fortiori*, macroscopique).

Il convient de noter que la typologie de ce niveau, contrairement à la section suivante (→ 4.2) est centrée autour d'options exclusives et absolument transversales : si une erreur est incohérente, alors elle n'est pas ambiguë et inversement, et toute erreur est soit incohérente, soit cohérente, soit ambiguë.

Après l'annotation de chaque erreur, nous obtenons la répartition suivante :

	Incohérentes	Cohérentes	Ambiguës
Nombre d'erreurs par type	185	131	83

Tab. 3 — Répartition des erreurs selon le niveau de cohérence

Comme nous pouvons le voir, les erreurs incohérentes sont les plus présentes, sans être majoritaires. Nous remarquons également une quantité non négligeable de cas d'ambiguïté dont il est intéressant d'étudier le fonctionnement : la sélection de la tête, dans des cas d'ambiguïté syntaxique, est-elle systématiquement liée à une minimisation de la longueur des dépendances (résolution au plus proche) ou le modèle est-il capable de proposer des solutions plus complexes ? Enfin, la quantité d'erreurs cohérentes dénote des difficultés de la prise en compte du contexte large et de la complexité à l'échelle de la phrase : une option peut sembler évidente dans un contexte de proximité

[31] Dans ce cas, nous pourrions parler de *variantes* puisque la prédiction n'est pas tout à fait erronée sur le plan syntaxique, mais une alternative à la prédiction est généralement hautement plus probable et des facteurs peuvent justifier ce choix.

tout en prévenant l'analyse syntaxique correcte de la phrase. Nous illustrons maintenant cette typologie.

4.1.1.1 Erreurs incohérentes

Les erreurs incohérentes, pour la majorité, peuvent être séparées en trois types : les incohérences syntaxiques liées au type de partie du discours, celles relevant plutôt de la morphologie et celles relevant déjà du plan syntaxique.

a) Incohérence de partie du discours

Il s'agit dans cette sous-section d'incohérences entre la prédiction et le type de partie du discours des tokens concernés. Certaines combinaisons sont en effet incompatibles : par exemple, un déterminant ne peut dépendre d'un verbe à un mode fini ; de même, un adverbe ne peut être sujet d'une proposition. TE-5 (fig. 10) illustre ce cas : le modèle prédit qu'un adverbe, *ici*, est sujet nominal (*nsubj*) du verbe *entretient*, ce qui est impossible et donc incohérent. De plus, le modèle prédit ici que *tout*, pronom dans ce cas[32], est déterminant d'un adverbe.

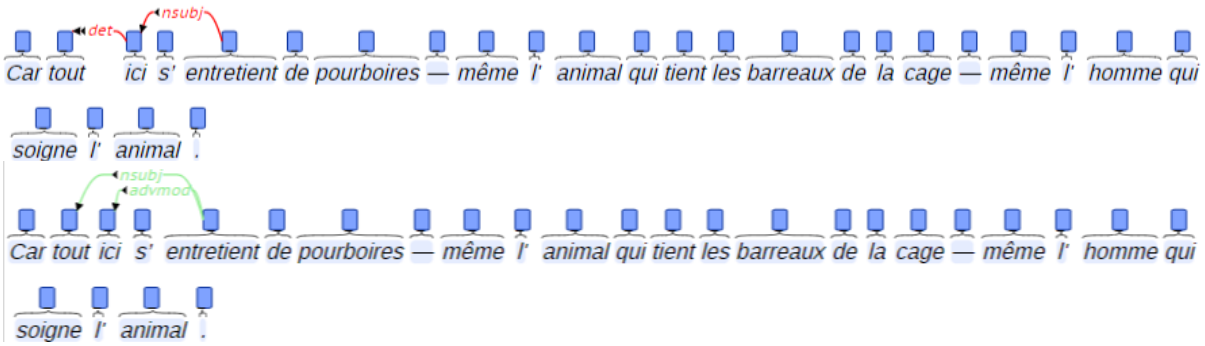


Fig. 10 — Prédiction et correction de TE-5

b) Incohérence morphologique

Il arrive également que la prédiction entre en contradiction avec des caractéristiques morphologiques des unités. Par exemple, dans C-6 (fig. 11), le parser prédit que *me* est le complément d'objet indirect du verbe *rassurent*, qui est pourtant transitif direct. Bien que *me* soit effectivement la forme de la première personne du singulier autant pour le complément d'objet direct que le complément d'objet indirect — ce qui peut conduire à des cas d'ambiguïté — il est impossible, sauf procédé rhétorique, qu'il s'agisse ici du complément d'objet direct.

[32] *Tout* peut également être adverbe et déterminant. Les cas d'homonymie constituent une difficulté pour le parser, comme cela est montré dans la section correspondante (→ 4.2.4).

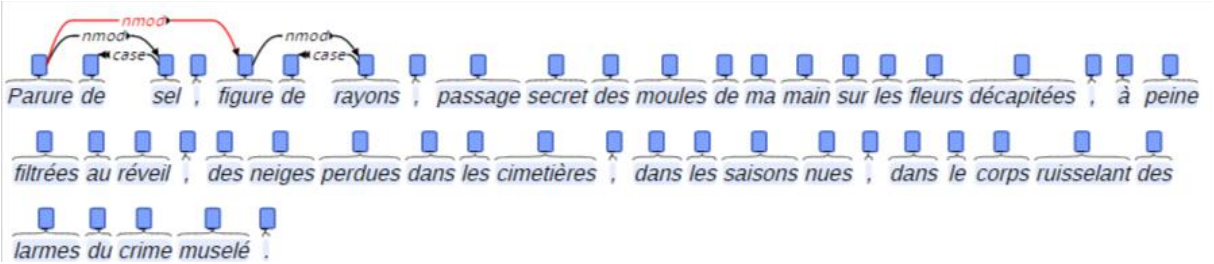
Fig. 11 — Prédiction et correction de C-6

En ce qui concerne EF-3 (fig. 12), c'est ici l'accord qui rend incohérente la prédiction : puisque *tracés* est accordé au masculin pluriel, il est impossible qu'il dépende de *meurtrissures*, substantif féminin pluriel. Il est donc plus plausible que son gouverneur soit *chemins*. Il s'agit d'ailleurs du seul candidat puisqu'aucun autre substantif masculin pluriel n'est présent dans la phrase.

Fig. 12 — Prédiction et correction d'EF-3

c) Incohérence syntaxique

L'incohérence peut dépendre également directement d'une construction syntaxique. Par exemple, dans R-3 (fig. 13), le modèle prédit que *figure de rayons* est un complément déterminatif de *Parure*, alors qu'il est en construction directe et n'intègre donc pas la préposition nécessaire à la construction du complément déterminatif. Il est plus probable qu'il s'agisse d'une apposition, bien que d'autres options comme la coordination entre *parure*, *figure* et *passage* sont également cohérentes en raison de l'absence de verbe.



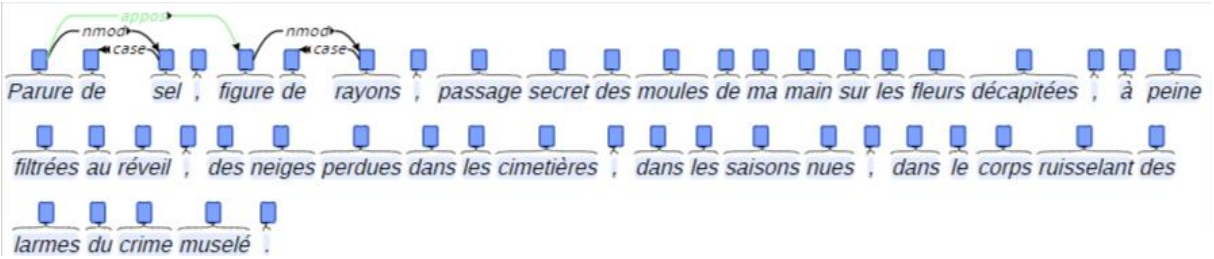


Fig. 13 — Prédiction et correction de R-3

4.1.1.2 Erreurs cohérentes

Les erreurs cohérentes sont extrêmement diverses et des types précis ne se dégagent pas réellement à l'intérieur de cette section ; nous illustrons cette catégorie par deux erreurs liées à des positions actancielles et une erreur liée à l'accord.

a) Actants macro-incohérents

L'exemple d'AR-1 (fig. 14) illustre un cas de surcharge actancielle : le verbe *aurait connu* reçoit deux dépendants étiquetés comme des compléments d'objet direct (*obj*) alors qu'il s'agit de fonctions soumises au « principe d'unicité », c'est-à-dire qu'une seule fonction de ce type (généralement actancielle) est autorisée par proposition. Ici, *quelqu'un qui n'en aurait pas l'air* semble plutôt être un cas de parataxe averbale. Cependant, puisque *quelqu'un* est compatible avec la fonction de complément d'objet direct (*Pierre a tué* ***quelqu'un***), nous estimons qu'il est *microcohérent*. C'est la présence d'un autre complément d'objet direct, *son cœur*¸ qui rend la prédiction macro-incohérente.

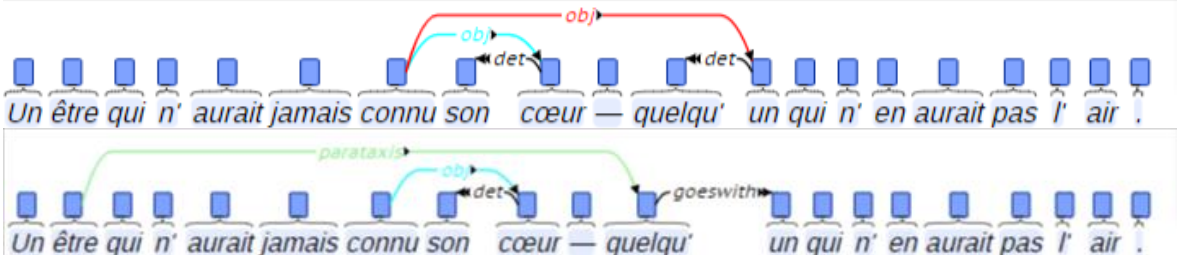


Fig. 14 — Prédiction et correction d'AR-1

Au contraire, la prédiction d'HE-7 (fig. 15) présente un déficit actanciel : dans la prédiction, la position sujet, qui est nécessaire lorsque le verbe est conjugué à un mode fini, n'est pas réalisée. Cela rend la prédiction *obl:mod* (complément adverbial non essentiel) macro-incohérente alors qu'elle est microcohérente puisqu'une proposition participe peut occuper ce rôle. Notons que des erreurs de segmentation interne à la phrase (→ 4.2.14.2) sont également présentes.

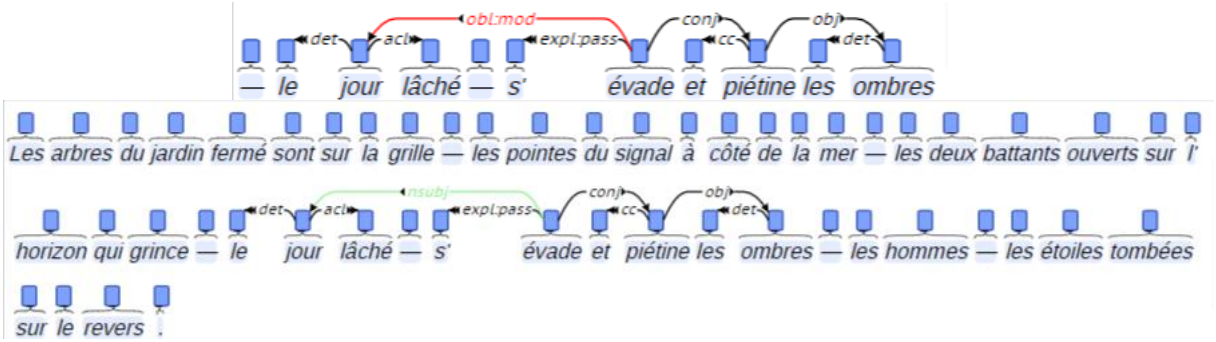


Fig. 15 — Prédiction et correction de HE-7

b) Accord macro-incohérent

APN-5 (fig. 16) présente un cas d'incohérence macro-syntaxique subtile : le syntagme prépositionnel *de ciel* est prédit complément déterminatif de *bâtiments*, ce qui est inhabituel au vu du nombre d'adjectifs intercalés, mais pour autant tout à fait cohérent. Cependant, le manque d'accord de *bleu*, alors qu'il est coordonné avec l'adjectif *rouges* au pluriel, indique, par les règles orthographiques d'accord des couleurs (Grevisse et Goosse, 2016 : 783), qu'il s'agit d'un nom de couleur accompagné d'un autre adjectif ou d'un substantif ; *de ciel*, seul candidat plausible à cette fonction, est donc nécessairement un dépendant de *bleu*.

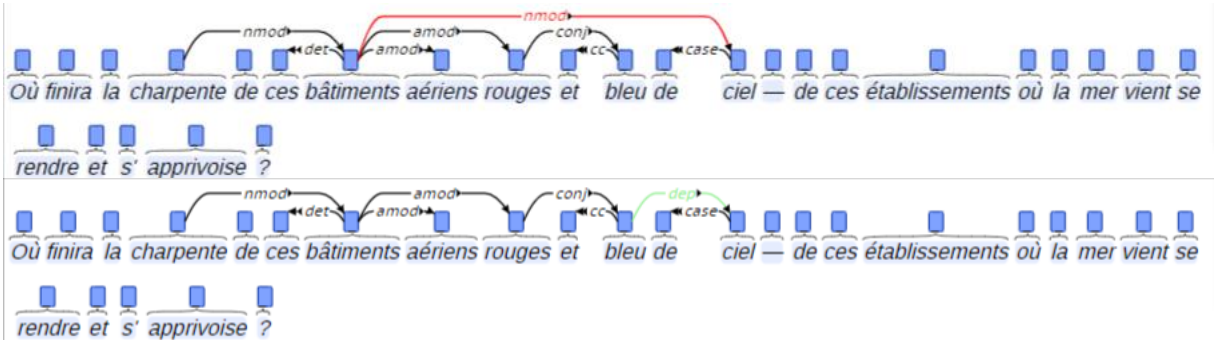


Fig. 16 — Prédiction et correction d'APN-5

4.1.1.3 Erreurs ambiguës

Les erreurs ambiguës peuvent apparaître dans la sélection du gouverneur, la prédiction de l'étiquette ou les deux. Comme nous l'avons déjà souligné, il s'agit plutôt de variantes ayant des niveaux différents de plausibilité sémantique que d'erreurs puisque la prédiction est cohérente sur le plan syntaxique seul.

a) Gouverneur (HEAD)

Dans le cas de FP-4 (fig. 17), il s'agit d'une variante de sélection du gouverneur pour le premier conjoint (*immobile*) d'une coordination occupant la fonction d'épithète. Deux options sont disponibles, *monde* et *paysan*, car il s'agit de deux substantifs au

masculin singulier. L'adjectif *immobile* est effectivement au singulier, mais son statut d'adjectif épicène nous prévient de déduire le genre du substantif auquel il se rapporte ; *muet*, coordonné avec lui, permet cependant d'indiquer que l'accord est au masculin. La morphologie ne nous permet donc pas de discréditer une option, et ce sont seulement les habitudes d'usage et le sens qui nous permet de supposer qu'il s'agit d'un *monde immobile et muet* plutôt que d'un *paysan immobile et muet*.

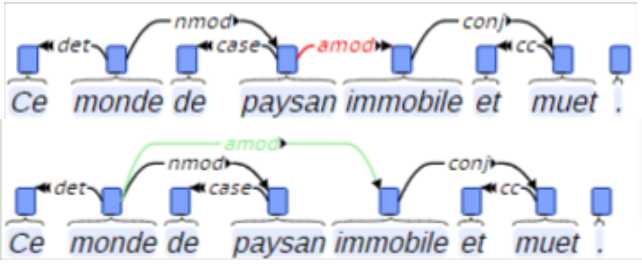


Fig. 17 — Prédiction et correction de FP-4

b) Étiquette (DEPREL)

Des cas similaires apparaissent au niveau de l'étiquette : dans le cas de BT-5 (fig. 18), le complément *à son filon* peut occuper la fonction de complément circonstanciel de lieu (*obl:mod*), indiquant le lieu de l'action, ou de complément d'objet indirect, indiquant la troisième entité impliquée dans l'action. *Donner* dispose de trois positions actancielles et l'analyse de *à son filon* comme complément d'objet indirect est cohérente puisque seules les positions sujet (*la lune*) et complément d'objet direct (*de l'or battu*) sont remplies. De même, un complément prépositionnel dépendant d'un verbe peut tout à fait occuper le rôle de complément adverbial non essentiel. La sémantique nous pousse à considérer cependant que la variante la plus plausible est celle du complément d'objet indirect antéposé.

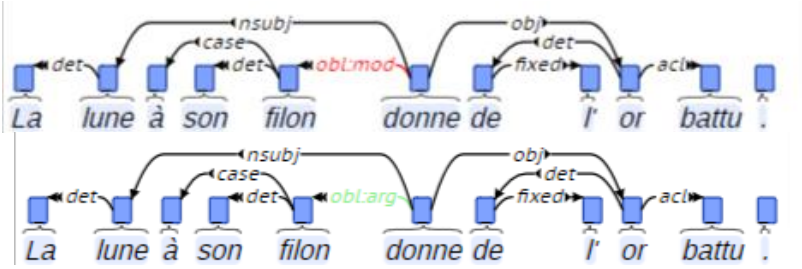


Fig. 18 — Prédiction et correction de BT-5

c) Gouverneur et étiquette (BOTH)

Enfin, des variantes plus radicales impliquant autant le gouverneur que l'étiquette peuvent survenir ; il s'agit du cas de CS-10 (fig. 19). Ici, la difficulté se situe au niveau

du syntagme prépositionnel *au temps*, prédit par le modèle comme complément déterminatif de *traînée*. Or, la position actancielle de complément d'objet indirect de *Mélangés* est disponible, bien que facultative, et *au temps qui se dévide* est le candidat idéal dans la phrase. Il s'agit donc de trancher entre le complément déterminatif dépendant de *trainée* (*la trainée au temps*) et le complément d'objet indirect de *Mélangés* (*mélangés au temps*). Puisque cette deuxième option est beaucoup plus plausible à nos yeux, il s'agit de la variante que nous avons sélectionnée. Cela implique donc une erreur BOTH.

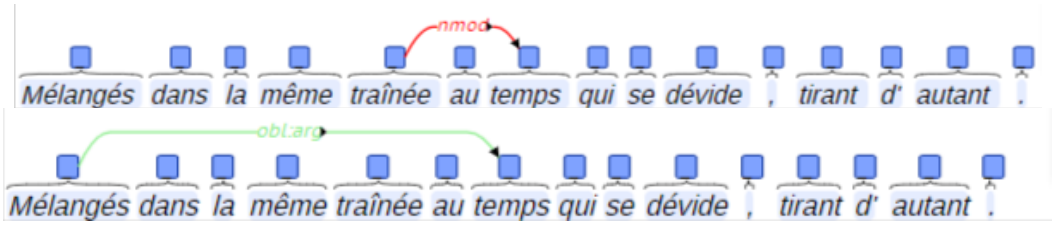


Fig. 19 — Prédiction et correction de CS-10

4.1.1.4 (In)cohérence et type d'erreur

Lorsque nous observons cet indicateur en fonction du type objectif (erreur d'étiquette, de gouverneur ou des deux), nous obtenons le nombre d'occurrences et la répartition suivante :

Erreurs	**HEAD**	**DEPREL**	**BOTH**
Incohérentes	13 %	48 %	39 %
Cohérentes	14 %	39 %	47 %
Ambiguës	35 %	39 %	26 %

Tab. 4 — Répartition des types d'erreurs selon le niveau de cohérence

Nous constatons tout d'abord une augmentation de fréquence importante des erreurs de gouverneur (HEAD) pour la catégorie des erreurs ambiguës par rapport aux autres. Cela peut s'expliquer par des cas dans lesquels plusieurs tokens sont des candidats crédibles pour la dépendance d'une fonction en particulier, mais que celui prédit par le modèle n'est pas optimal. Ensuite, nous pouvons noter une forte représentation des erreurs d'étiquette pour les erreurs incohérentes. Cela n'est pas surprenant : l'incohérence comme nous l'avons définie vient généralement de l'assignation à un token d'une étiquette incompatible avec le couple gouverneur-dépendant donné. Dans ce cas, des erreurs HEAD nous semblent plutôt être le fruit du hasard, puisque le modèle doit prédire une étiquette incompatible uniquement avec le gouverneur — puisque seul le gouverneur doit être modifié pour la correction — alors qu'il existe un

candidat crédible à ce poste. Enfin, la surreprésentation de la catégorie BOTH dans le cas des erreurs cohérentes est plus difficilement interprétable.

4.1.2 Paramètres quantitatifs de complexité linguistique

Après cette présentation des niveaux de cohérence, nous analysons divers paramètres quantitatifs pouvant être importants dans l'apparition d'erreurs d'analyse syntaxique. En effet, dans leurs travaux, Liu (2008) et Yan et Kahane (2018) proposent des indicateurs permettant d'estimer la complexité linguistique d'une phrase. Nous les présentons ici et observons leur variation dans le corpus. Il s'agit de la moyenne de la longueur des dépendances (→ 4.1.2.1) et de la taille, du poids et du poids combiné du flux de dépendance (→ 4.1.2.2).

4.1.2.1 Moyenne de la longueur des dépendances

Un facteur lié à la longueur de phrase (Jiang et Liu, 2015 : 103) et proposé par Liu (2008) comme indicateur de la complexité linguistique ou de la difficulté de la compréhension du langage est la *moyenne de la longueur des dépendances* (MLD). Il s'agit d'un paramètre indiquant, dans une phrase ou un corpus, le nombre de tokens intercalés[33] entre un gouverneur et son dépendant : pour obtenir la longueur d'une dépendance, il suffit de calculer la valeur absolue de la différence entre l'index des deux tokens concernés ($LongDep = |idx_{tok1} - idx_{tok2}|$). Ainsi, dans l'exemple suivant (LPN-1), la longueur de la dépendance entre *tombe* (racine de la phrase) et *Doucement* (complément adverbial, *advcl* dépendant de *tombe*) est de 9 (|10-1)), car les tokens 2 à 9 sont intercalés ; celle entre *poudre* et *saveur* est de 2 (|3-5|), car le token 4, *de*, les sépare.

(LPN-1) Doucement$_1$ la$_2$ poudre$_3$ de$_4$ saveur$_5$,$_6$ couleur$_7$ des$_8$ fonds$_9$ tombe$_{10}$ sur$_{11}$ la$_{12}$ moulure$_{13}$ égratignée$_{14}$ du$_{15}$ cadre$_{16}$.$_{17}$

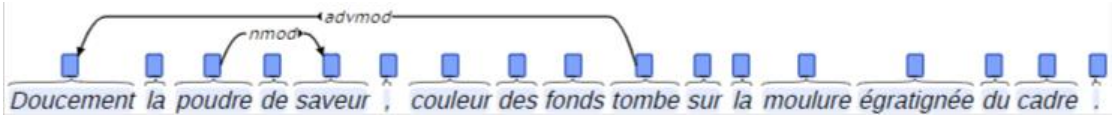


Fig. 20 — Dépendances de LPN-1 de longueur 9 et 2

En faisant la moyenne des *n*-1 dépendances (puisque la dépendance de la racine, qui pointe vers l'index 0 dans UD, est ignorée), nous obtenons la MLD. Comme nous

[33] L'indicateur ne correspond pas exactement au nombre de tokens intercalés, mais à ce nombre augmenté de 1. En effet, deux tokens adjacents dont l'un dépend de l'autre ont une longueur de dépendance de 1.

l'avons évoqué, bien que Liu (2015) montre le lien entre longueur de la phrase et moyenne de la longueur des dépendances, ces deux indicateurs ne sont pas équivalents et la MLD est beaucoup plus précise : dans la figure 21, bien que les deux phrases soient d'une longueur parfaitement équivalente, la moyenne de la longueur des dépendances est de 1,9 pour la première et 3 pour la seconde.

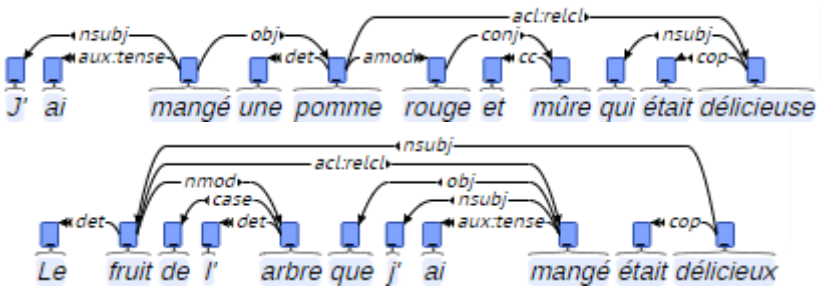


Fig. 21 — Phrases à 11 tokens, mais de MLD différente

La seconde phrase, de MLD plus élevée, semble plus compliquée à annoter que la première, avec des dépendances plutôt longues et une ambiguïté syntaxique concernant l'antécédent de *que*, qui pourrait se rapporter à *fruit* ou à *arbre*, ambiguïté qui disparaît si on supprime ou déplace les tokens intercalés entre *fruit* et *que*. La longueur des dépendances est en effet une difficulté, en particulier pour les parsers basés sur des mécanismes de transition comme le nôtre :

> Many, if not most, of the current state-of-the-art parsing systems are based on this framework [les mécanismes de transition], and all the different algorithm variants that are at its core have in common that they build short dependencies before (and requiring fewer transitions than) long ones (Gómez-Rodríguez, 2017 : 102)

La minimisation de la longueur moyenne des dépendances est, comme le montre Liu (2007) un universel du langage lié à la grammaire et à des mécanismes cognitifs pour garantir l'intelligibilité.

4.1.2.2 Flux de dépendance : taille, poids et poids combiné

Un autre concept théorique susceptible d'aider à la définition de critères objectifs et quantitatifs de complexité syntaxique est celui de *flux de dépendance*. Selon Kahane, Yan et Botalla (2017 : 74), le flux de dépendance se définit comme l'ensemble des dépendances qui, dans une position entre deux tokens, lient un mot à gauche de cette position avec un mot à droite de cette position (fig. 22). Ainsi, le flux entre *s'* et *étire* est constitué des dépendances « *campagne* <nsubj *étire* » et « *s'* <expl:pass *étire* ».

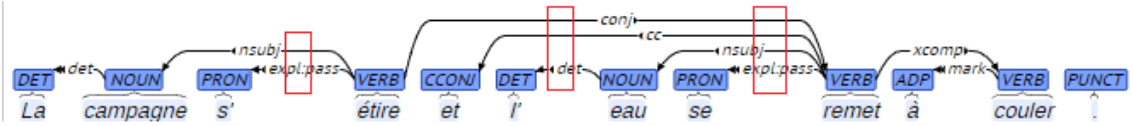


Fig. 22 — Flux de dépendance à trois positions différentes

Nous présentons les concepts de *taille du flux*, *poids du flux* et *poids combiné du flux*.

a) Taille du flux

La taille du flux est le nombre de ces dépendances. Puisque le flux est considéré entre deux mots, il s'agit d'une donnée ponctuelle : le flux et sa taille varient en fonction de la position choisie. Kahane, Yan et Botalla (2017) fournissent l'exemple suivant :

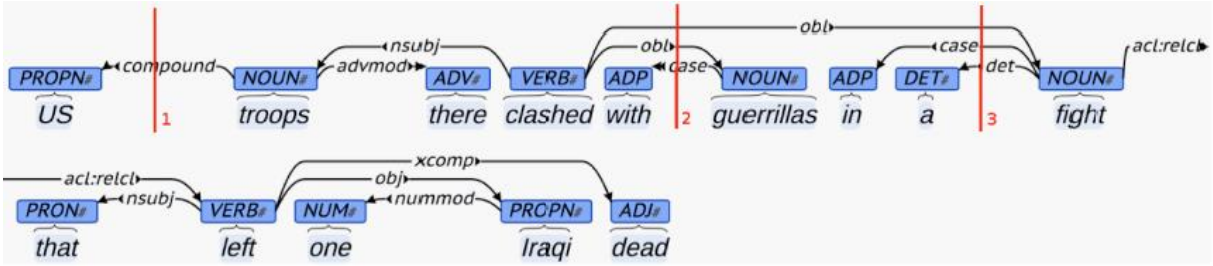


Fig. 23 — Arbre syntaxique UD avec des positions entre les mots représentées (Kahane, Yan et Botalla, 2017 : 74)

Dans cet exemple (fig. 23), à la position marquée 1, le flux est constitué de l'unique dépendance entre *troops* à droite et *US* à gauche, étiquetée *compound* (*US* <compound *troops*). Sa taille est donc de 1. En position 3, il y a trois dépendances liant des mots de droite et des mots de gauche : « *clashed* obl> *fight* », « *in* <case *fight* » et « *a* <det *fight* ». Sa taille est donc de 3. La taille de ce flux a des implications sur la cognition des locuteurs : « *The flux represents the set of pending syntactic relations that the speaker has to keep in mind after every word* » (Kahane, Yan et Botalla, 2017 : 74).

b) Poids du flux

Le poids du flux de dépendance équivaut plus ou moins à la profondeur (le nombre de niveaux) que nous rencontrerions dans une analyse en constituants immédiats, donc le nombre de segments centrés autour d'une tête (Kahane, Yan et Botalla, 2017 : 75). Il est égal à la taille du plus grand sous-ensemble du flux ne contenant que des dépendances disjointes (c'est-à-dire ne partageant aucun nœud ; fig. 24). En pratique, il s'agit de regrouper dans un sous-ensemble le plus de dépendances possibles sans jamais qu'un token ne soit utilisé deux fois.

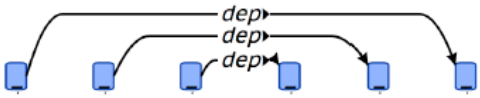


Figure 3. Disjoint dependencies

Fig. 24 — Dépendances disjointes (Kahane, Yan et Botalla, 2017 : 74)

Par exemple, à la position 2 (fig. 23), le plus grand sous-ensemble est de taille 2 (« *clashed* obl> *fight* » et « *with* <case *guerrillas* ») ; le poids du flux à la position 2 est donc de 2. À la position 3, il y a plusieurs sous-ensembles à 1 dépendance, dont la plus longue est « *clashed* obl> *fight* » ; le poids du flux est donc de 1.

c) Poids combiné du flux

En combinant longueur des dépendances et poids du flux, Yan et Kahane (2018) proposent un nouvel indicateur de complexité du traitement de la phrase : le poids combiné (*combined weight*) :

> Our hypothesis of the complexity for sentence processing considers two aspects. On the one hand, the complexity depends on the number of disjoint dependencies that we measure by flux weight; on the other hand, it depends on the dependency length (modulo granularity). Thus, by combining the length of dependencies and the flux weight, we introduce a new measure, which we call the combined weight of the flux. […] The combined weight in a given position is the sum of the dependency length of the longest disjoint dependencies. (Yan et Kahane, 2018 : 40-41)

Le poids combiné à la position 2 est de 6, addition de la taille des dépendances du plus grand sous-ensemble de dépendances disjointes (5, longueur de « *clashed* obl> *fight* » et 1, longueur de « *with* <case *guerrillas* »). Le poids combiné à la position 3 est 5.

4.1.2.3 Application à nos données

Observons ces indicateurs à l'intérieur de nos données. Le tableau ci-dessous envisage les différents niveaux de cohérence en fonction des paramètres quantitatifs présentés précédemment : la longueur des dépendances, la taille, le poids et le poids combiné du flux. Afin d'isoler au maximum les erreurs du reste du corpus, nous avons décidé d'envisager le flux à proximité des erreurs, c'est-à-dire dans les positions directement à gauche et à droite de l'erreur ; nous en faisons ensuite la moyenne. Concernant la longueur, seule la longueur des dépendances des tokens erronés est considérée. Le tableau présente ces paramètres dans le corpus original, dans le corpus corrigé ainsi que la variation entre les deux en pour cent. Nous avons décidé d'inclure

également les erreurs DEPREL, dont la correction n'a pourtant aucune influence sur le flux (puisqu'il ne s'agit que d'erreurs d'étiquette), afin de pouvoir comparer dans un second temps ces paramètres à proximité des erreurs par rapport à la moyenne du corpus.

	Longueur de la dépendance (O|C)		**Taille du flux à proximité (O|C)**		**Poids du flux à proximité (O|C)**		**Poids combiné du flux à proximité (O|C)**	
Incohérentes	6,55	6,67	2,37	2,48	1,23	1,30	11,67	13,16
	+2 %		+4 %		+6 %		+13 %	
Cohérentes	4,31	5,31	2,12	2,17	1,16	1,22	9,78	11,74
	+23 %		+3 %		+5 %		+20 %	
Ambiguës	5,54	5,78	1,97	2,10	1,16	1,21	10,11	11,17
	+4 %		+7 %		+5 %		+11 %	

Tab. 5 — Évolution des paramètres quantitatifs selon les différents types

Nous remarquons particulièrement une tendance à la hausse des valeurs après la correction : la variation entre le corpus original et le corpus corrigé est positive pour la totalité des paramètres, sans aucune exception. Puisque nous avons identifié que chacun de ces paramètres est en puissance un facteur de complexité linguistique, nous pouvons formuler l'hypothèse que le modèle a tendance à minimiser la complexité lors de la prédiction. Ceci est particulièrement visible pour le poids combiné du flux, qui semble se démarquer parmi ces indicateurs et pourrait confirmer son intérêt dans l'analyse de la complexité syntaxique. Notons au contraire que la variation de longueur des dépendances est assez faible — à l'exception des erreurs cohérentes — contrairement à ce que nous aurions pu attendre : le modèle ne semble pas systématiquement prédire comme gouverneur le candidat le plus proche pour la résolution de l'ambiguïté, ce qui est particulièrement important à noter puisque cela montre une plus grande subtilité de la prédiction.

Enfin, lorsque nous envisageons la moyenne du corpus par rapport au contexte proche des erreurs (tab. 6), nous pouvons constater que les valeurs de deux indicateurs (la longueur des dépendances et le poids combiné) sont considérablement plus grandes à proximité des erreurs.

Nous formulons donc les hypothèses suivantes : **1.** le modèle a davantage tendance à réaliser des erreurs de prédiction lorsque la longueur des dépendances est élevée et que le poids combiné du flux l'est aussi ; **2.** le modèle réalise des erreurs qui minimisent

en moyenne les indicateurs, faiblement dans la plupart des cas, mais cela est assez marqué pour le poids combiné. Ces indicateurs nous semblent importants et il serait intéressant de conduire une analyse plus complète les concernant, mais cela n'est pas l'objet de ce travail.

	Longueur des dépendances	**Taille du flux**	**Poids du flux**	**Poids combiné du flux**
Corpus entier	3,90	2,14	1,26	7,99
Variation par rapport aux positions adjacentes aux erreurs[34]				
Incohérentes	+71 %	+16 %	+3 %	+65 %
Cohérentes	+36 %	+2 %	-3 %	+47 %
Ambiguës	+48 %	-2 %	-4 %	+40 %

Tab. 6 — Paramètres quantitatifs du corpus corrigé entier et variation par rapport aux erreurs

Il s'agit maintenant de nous intéresser de plus près aux erreurs et d'en proposer une typologie.

4.2 Phénomènes linguistiques et structures syntaxiques : vers une typologie plus précise des erreurs

Dans cette section, nous classons les erreurs rencontrées par le parser selon des structures (comparaison relative, phrases clivées…), fonctions syntaxiques (*xcomp*, apposition…) ou phénomènes morphosyntaxiques (figement, ellipse…). Dans chaque section, nous illustrons le type d'erreur représenté et tentons d'expliquer le comportement du modèle face à ces situations.

L'ordre de la présentation est arbitraire. La majorité de l'exposé est organisé selon la fréquence des erreurs considérées dans chacune des sections. Cependant, nous distinguons, et rejetons en fin d'exposé **1.** les erreurs qui nous semblent fortement liées au modèle d'annotation UD (→ 4.2.12) ; **2.** les structures uniques et exceptionnelles, ne permettant aucune conclusion solide (→ 4.2.13) ; **3.** les erreurs liées à des paramètres techniques et configurations du parser, qui auraient pu ne pas apparaître en raison d'une configuration différente des paramètres de l'analyse (→ 4.2.14).

[34] Une augmentation signifie que la valeur de l'indicateur pour ce type d'erreur est plus élevée que la moyenne du corpus ($Pourcentage = \frac{Valeur_{Erreur} - Valeur_{Corpus}}{Valeur_{Corpus}}$)

4.2.1 Relations d'équivalence et d'entassement

Cette section concerne les relations d'équivalence et d'entassement (Rossi-Gensane, 2017), phénomènes également désignés dans le langage oral sous le nom de « piles » ou « listes », incluant la coordination, l'apposition, la reformulation et la disfluence (Lacheret-Dujour *et al.*, 2019 : 49). Après diverses considérations théoriques, nous expliquons la posture adoptée durant la correction et présentons les erreurs.

a) Une difficulté théorique

Ces relations *horizontales* (c'est-à-dire qu'elles n'établissent pas une hiérarchie entre deux mots) sont un problème récurrent de la grammaire dépendancielle puisqu'il ne s'agit pas de relations de dépendance. Dès Tesnière et la « jonction » (1959 : 323 ; fig. 25), ligne horizontale contrairement aux connexions qui ont toujours une orientation verticale afin de marquer l'asymétrie entre le gouverneur et le dépendant, des solutions sont proposées pour la description de ces structures problématiques.

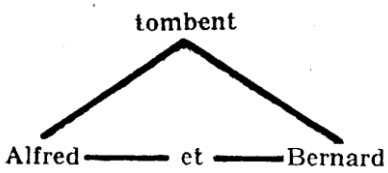


Fig. 25 — Jonction chez Tesnière (1959 : 327)

En effet, les éléments d'une pile entretiennent une relation particulière sur le plan paradigmatique introduisant des difficultés de description en dépendance, en particulier pour la coordination qui peut être décrite d'une façon « symétrique » ou « asymétrique » :

> [a pile is a] list of elements [that] piles up in the same syntactic position. We therefore group the analysis of coordination together with the analysis of other phenomena such as reformulation, disfluency, partial answer or negotiation. The elements of a pile are linked to one another by a relation that is both syntagmatic (they follow one another) and paradigmatic (they fill the same syntactic slot with respect to their common governor). (Lacheret-Dujour *et al.*, 2019 : 69)
>
> When an element piles up with another element, it has a special relation – which Blanche-Benveniste would call a *paradigmatic relation* – with the first element, and is indirectly governed by the first element's governor. This triangular relationship cannot be easily rendered in pure government structures. […] Syntactic analyses of coordination can generally be divided into two families : symmetrical and asymmetrical analyses. […] Symmetrical analyses aim to give equal status to each conjunct, disregarding the linear order. Asymmetrical analyses on the contrary give a

> special status to one, commonly the first, of the conjuncts, and iteratively place the other conjuncts below the special one. (Lacheret-Dujour *et al.*, 2019 : 72)

Le *bubble tree* (arbre à bulles) (Kahane, 1997) est un exemple d'analyse symétrique de la coordination cohérente du point de vue mathématique. Au contraire, les descriptions inspirées de Mel'čuk (Kahane, 2001 : 6), dont fait partie UD, ont opté pour la description asymétrique : les relations d'équivalence et d'entassement sont donc décrites comme un arbre de dépendance dont la tête est le premier conjoint. Notons que, dans le cas de la coordination, comme le relèvent Gerdes *et al.* (2019), des différences théoriques sont encore notables entre les systèmes d'annotation comme UD dans lesquels les conjoints sont organisés en bouquet (c'est-à-dire que tous les conjoints à partir du second dépendent du premier) de ceux comme SUD dans lesquels les conjoints sont organisés en chaînes (c'est-à-dire que tous les conjoints à partir du second dépendent du précédent) ; cela est illustré à la figure 26.

Fig. 26 — Coordination en bouquet et en chaîne

Cette solution asymétrique, pour le mérite qu'elle a d'être applicable dans une perspective purement dépendancielle de l'annotation syntaxique, ne permet pas de rendre compte du fait que les éléments « se trouvent dans un même type de rapport vis-à-vis de leur point d'incidence » (Feuillard-Aymard, 1989 : 122, cité depuis Rossi-Gensane, 2017 : 65) et que « les éléments qui se trouvent [en relation d'équivalence] impliquent la mise en facteur commun de l'élément auquel ils se rattachent ; […] ils se situent à un même niveau hiérarchique » (Feuillard-Aymard, 1989 : 137, cité depuis Rossi-Gensane, 2017 : 69). De plus, ce phénomène est à distinguer de celui de coexistence, pour lesquels le *principe d'unicité* est déterminant :

> Enfin, Feuillard-Aymard (1989) distingue la coordination de la simple coexistence, qui concerne uniquement des éléments assumant des fonctions non soumises au principe, parfois dit d'unicité, d'une seule fonction par proposition, tels, par exemple, les compléments circonstanciels. (Rossi-Gensane, 2017 : 72)

Nous considérons également comme relevant de cette section l'apposition, la parataxe et la reprise syntaxique, bien que cette première ne fasse pas l'objet d'un consensus :

> L'apposition ne semble pas rangée de manière aussi consensuelle que la coordination dans les relations d'équivalence et dans les relations d'entassement. Il s'agit d'ailleurs, plutôt que d'une relation, d'une fonction, incidence à un nom ici et qui ne saurait donc partager avec ce dernier un même point d'incidence. Ainsi définie, l'apposition ne peut être vue comme exerçant (parallèlement) la même fonction que son point d'incidence, mais seulement comme ayant la possibilité d'en prendre la relève (on glisse alors vers une même fonction potentielle). (Rossi-Gensane, 2017 : 77)

La reprise syntaxique est distinguée de l'apposition, ce qui est également marqué dans la structure de l'annotation dans UD (→ 4.2.1.4) :

> La reprise (syntaxique) […], en particulier, note le détachement avec rappel pronominal. Dans ce cas, l'élément détaché est un syntagme susceptible de se trouver à gauche ou à droite d'un élément pronominal, caractérisé à l'oral par une prosodie particulière et isolé à l'écrit du reste de la phrase par une ponctuation faible comme la virgule. Neveu (2000b, p. 113) évoque à ce propos les « détachements par redoublement d'actant […] du type topique » et insiste sur le côté référentiel des topiques (que l'on peut définir comme ce à propos de quoi l'on parle) « qui les distingue […] des appositions ». Les topiques sont ainsi coréférentiels des éléments pronominaux qu'ils reprennent […]. (Rossi-Gensane, 2017 : 91)

Dans le cas de la parataxe, type particulier de coordination entre deux propositions, il ne s'agit en réalité pas d'une pile :

> Kahane & Pietrandrea (2012, p. 1813) précisent que « n'est pas considéré[e] comme [un] entassement la coordination reliant deux ou plusieurs unités rectionnelles » puisque, « par définition, les unités rectionnelles sont des unités maximales pour la rection et [que,] par conséquent, […] deux unités rectionnelles coordonnées [ne sauraient occuper] une même position régie, puisque position régie il n'y a pas ». (Rossi-Gensane, 2017 : 74)

Cependant, il est de même incohérent de considérer qu'une des deux propositions dépend de l'autre dans ce cas ; la relation les unissant est donc au moins *horizontale*, ce qui justifie leur présence dans cette section.

Il est amusant de constater que les erreurs de segmentation interne (→ 4.2.14.2) semblent justifier empiriquement notre choix de traiter dans une même section la coordination, l'apposition, la parataxe et la reprise syntaxique : il s'agit des seuls types de relations (ou fonctions) ayant pu mener à une division incorrecte interne à la phrase par le parser.

b) Position adoptée dans la différenciation des relations

Il est parfois difficile de différencier les relations horizontales. Nous avons donc essayé de définir une série de critères, objectifs dans la mesure du possible, permettant une correction cohérente des résultats du parsing.

Tout d'abord, la distinction entre éléments coordonnés et éléments en situation de coexistence est liée à la présence d'une conjonction de coordination, à la présence de dépendants co-référents — c'est-à-dire des structures syntaxiques dépendant de plusieurs conjoints en même temps[35] — et au principe d'unicité. En effet, nous considérons qu'il y a coordination si **1.** une conjonction de coordination est présente, même si elle n'est explicite qu'entre les deux derniers conjoints d'une coordination à trois ou plus éléments (*Victor, David et moi*) ; **2.** les éléments possèdent un dépendant co-référent en commun (voir note 35) ; **3.** il s'agit d'une position soumise au principe d'unicité (une seule position de ce type par proposition), généralement dans le cas d'actants[36]. Ces critères sont également applicables pour différencier la coordination de la parataxe lorsqu'il s'agit de propositions. Notons que le corpus contient de nombreux cas de parataxe incluant une proposition averbale en raison des nombreuses ellipses du verbe. La différence entre coordination et coexistence est importante puisque, en plus de l'étiquette syntaxique, le gouverneur varie (fig. 27) : dans le cas d'une coordination, le premier conjoint dépend du gouverneur de la coordination et les conjoints suivants dépendent de ce premier conjoint, alors que dans le cas d'une coexistence, tous les éléments dépendent du même gouverneur. Les structures syntaxiques ne sont donc pas superposables dans UD.

Fig. 27 — Coexistence et coordination de compléments déterminatifs

Dans le cas de l'apposition et de la reprise syntaxique, nous considérons qu'il s'agit d'une reprise syntaxique si et seulement si l'élément le plus proche du verbe est un

[35] Puisque la structure mathématique reste un arbre, les dépendants co-référents ne dépendent pas explicitement des deux conjoints, mais une relation triangulaire similaire à celle entre les conjoints et leur gouverneur est implicite. Par exemple, dans la phrase *sur le tronc, au sommet, au pied de l'arbre, les écureuils s'amusent*, nous pouvons affirmer que *tronc*, *sommet* et *pied* sont en situation de coordination et pas de coexistence, car le complément déterminatif *de l'arbre* se réfère aux trois substantifs.

[36] Dans *J'achète du sucre, de l'orge*, *sucre* et *orge* sont coordonnés, car il s'agit de compléments d'objet directs, position soumise au principe d'unicité.

pronom seul (il ne peut pas être déterminé par une proposition relative par exemple). Les structures ne sont pas superposables non plus dans UD à la position sujet, puisque le sujet, dans le cas de l'apposition, est le premier des deux éléments (***Le chat***, *Félix, mange une souris*) alors qu'il s'agit du second dans le cas de la reprise syntaxique (*Pierre, **il** mange une pomme*). La relation entre les éléments en situation horizontale est généralement vers l'avant, le futur de la phrase considérée dans sa linéarité (« *chat* appos> *Félix* ») dans le cas de l'apposition, mais régulièrement vers l'arrière, le passé de la phrase (« *Pierre* <dislocated *il* ») dans le cas de la reprise syntaxique.

Enfin, la distinction entre apposition et coordination est particulièrement compliquée dans notre corpus en raison de sa complexité sémantique. Il est en effet parfois difficile de savoir si les différents groupes nominaux désignent tous un même référent qui est reformulé ou redéterminé (*le chat, ce félin, Félix* ; plutôt apposition) ou si les référents sont différents (*le chien, le tigre, l'ours* ; plutôt coordination). Il s'agit donc d'une analyse au cas par cas. Notons que cela est parfois explicite : lorsqu'il s'agit de la position sujet, l'accord du verbe peut rendre explicite la structure. En effet, dans *le chat, Félix, est dans le jardin*, il s'agit d'une apposition puisque le verbe est au singulier, cela signifie que le sujet est au singulier. Au contraire dans *le chat, Félix, sont dans le jardin*, il s'agit d'une coordination puisque *le chat* **et** *Félix* se trouvent dans le jardin. Cependant, dans ce cas, les structures syntaxiques sont superposables.

4.2.1.1 Coordination et coexistence

Le parser a réalisé **35 erreurs**[37] réparties en 22 phrases[38] liées à la coordination ainsi que **37** liées à la coexistence, en 18 phrases[39]. Il s'agit généralement d'erreurs impliquant au moins la prédiction du gouverneur, régulièrement ambiguës. Bien que les erreurs soient très diverses, trois types se dégagent : l'erreur d'identification des conjoints, l'omission d'une coordination et la confusion entre coordination et coexistence.

[37] Ce nombre inclus les erreurs (10) liées à la fois à la coordination ET à la coexistence, c'est-à-dire les situations dans lesquelles le parser a prédit une coordination alors qu'il s'agit d'une coexistence ou inversement.

[38] PT-2, PT-3, G-7, VF-5, SA-5, SA-6, DC-5, APN-8, RI-4, BT-2, LPN-3, LPN-4, LPN-7, TN-2, HE-4, HE-5, HE-6, NV-8, GL-2, GL-7, EF-1 et TPB-7.

[39] FP-1, CS-11, SA-6, DC-5, APN-6, R-3, BT-2, TN-2, AA-4, AP-5, AP-13, GL-6, EF-4, TPB-1, TPB-4, TPB-5, TPB-14 et TPB-17.

a) Erreur d'identification des conjoints

Dans un certain nombre de cas, il arrive que les conjoints prédits par le modèle ne soient pas corrects parce que le modèle sélectionne deux conjoints qui ne sont pas situés à un même niveau syntagmatique et ne peuvent donc pas être coordonnés. Il est également courant, lorsque plusieurs conjonctions de coordination sont présentes dans une structure à trois conjoints ou plus, par exemple dans des coordinations corrélatives, que les conjoints soient appairés deux à deux avec le conjoint précédent. La phrase G-7 (fig. 28) en fournit un exemple : pour la coordination à trois éléments (*nuit*, *homme* et *Dieu*), dans laquelle la conjonction de coordination *ni* est répétée, le modèle prédit une structure en chaîne plutôt qu'une structure en bouquet comme cela est de mise dans l'environnement UD. La structure n'est donc pas reconnue correctement, puisque dans ce cas, plutôt que de coordonner les trois conjoints d'un seul mouvement (*nuit — homme — Dieu*), un élément unique est d'abord formé avec les deux derniers conjoints, avant qu'il soit coordonné avec le premier (*nuit* – [*homme* – *Dieu*]). Nous pouvons supposer que cette erreur est due au fait que le modèle a pour habitude de coordonner ensemble deux éléments adjacents à une conjonction de coordination, sans considérer le contexte plus en profondeur. Cela montre que les éléments très proches d'une position sont beaucoup plus déterminants pour la prédiction que le contexte de la phrase.

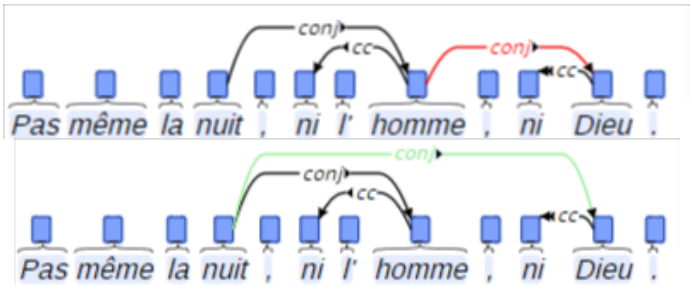


Fig. 28 — Prédiction et correction de G-7

b) Omission d'une coordination

Il est également courant que le modèle ne reconnaisse tout simplement pas que des éléments sont coordonnés lorsque la conjonction de coordination n'est pas explicite. Dans le cas de SA-6 (fig. 29), la coordination entre les trois propositions, dont le verbe est implicite pour les deux dernières (*les têtes remuées en rond près des rideaux* et *les visages éclairés contre la ville*), n'est pas entièrement reconnue. En effet, seule la coordination entre les deux dernières est prédite, ce qui est probablement dû au fait que la conjonction de coordination, *et*, soit explicite. Or, il est d'usage lors d'une

coordination de plus de deux conjoints de placer la conjonction entre les deux derniers conjoints ; elle est donc implicite aux autres positions.

Dans ce cas, la coordination est probablement plus difficile à prédire en raison de l'ellipse du verbe, puisqu'il est inhabituel pour le modèle de prédire la relation *conj* pour des relations entre des éléments n'ayant pas la même nature. L'absence de conjonction explicite a donc dû être déterminante.

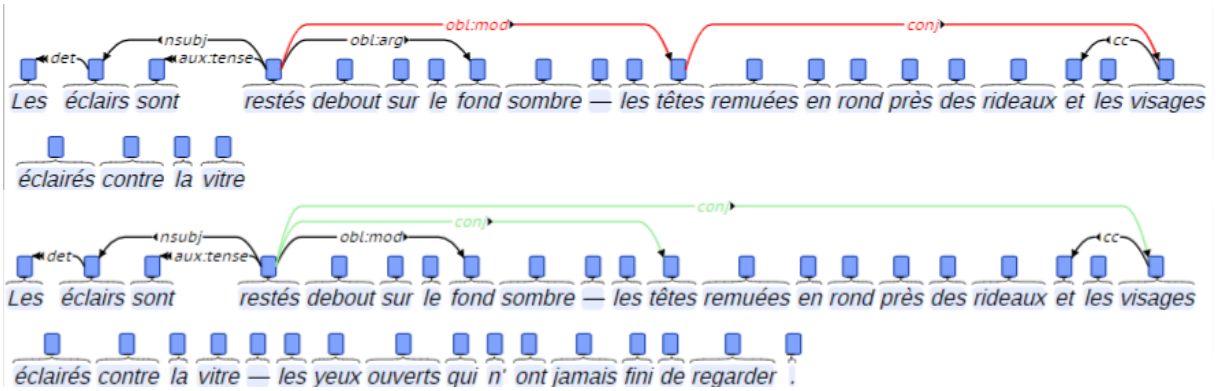


Fig. 29 — Prédiction et correction de SA-6

Notons que cette erreur n'est pas du tout systématique et que le modèle prédit correctement les coordinations à trois conjoints à plusieurs reprises dans notre corpus. Dans le cas de TPB-5, le modèle prédit même correctement une coordination à sept conjoints.

> (TPB-5) Ô toi qui traines sur la vie, entre les buissons fleuris et pleins d'épines de la vie, parmi les feuilles mortes, les reliefs de triomphes, les appels sans secours, les balayures mordorées, la poudre sèche des espoirs, les braises noircies de la gloire, **et** les coups de révolte, toi, qui ne voudrais plus désormais aboutir nulle part.

c) Confusion entre coordination et coexistence

Au contraire, le modèle a parfois considéré la présence d'une coordination entre plusieurs éléments alors qu'ils sont en réalité en situation de coexistence. Il s'agit par exemple du cas d'adjectifs épithètes en situation de pile dans TPB-1 (fig. 30). Dans ce cas, en l'absence de conjonction de coordination, de co-référents et l'épithète n'étant pas soumise au principe d'unicité, les adjectifs *doré*$_1$, *rouge*, *glacé* et *doré*$_2$ coexistent et occupent une pile paradigmatique épithète d'*abîme*.

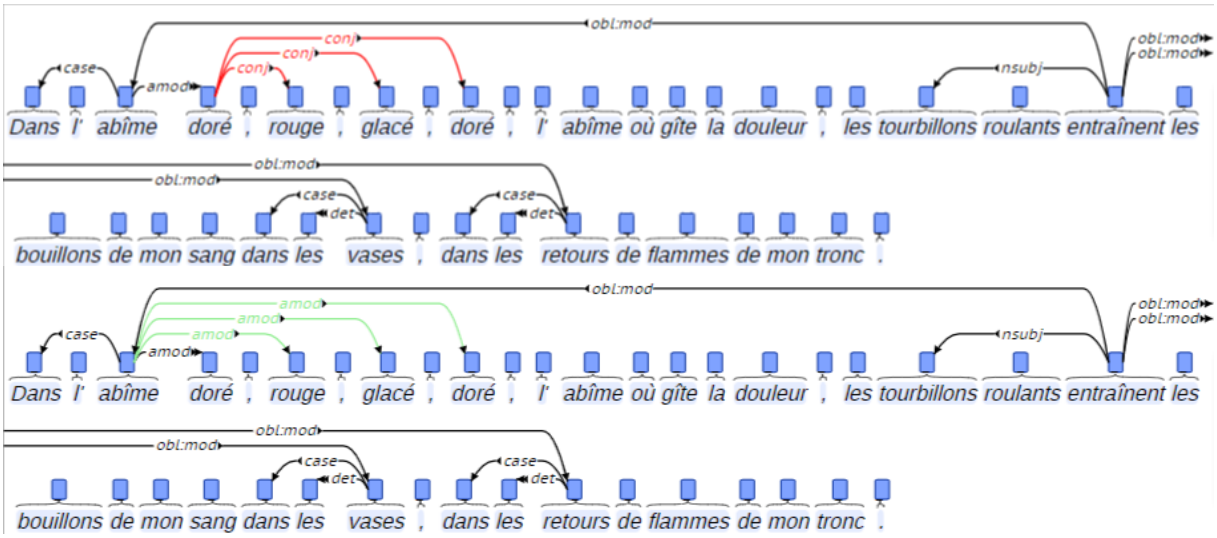


Fig. 30 — Prédiction et correction de TPB-1

4.2.1.2 Parataxe

Nous rencontrons **13 erreurs** liées à la parataxe, souvent dans des cas de faux négatif. Certaines sont ponctuelles, mais nous pouvons remarquer deux groupes : les erreurs influencées par une ellipse du verbe ainsi que la substitution par l'étiquette *advcl*.

Il est régulier de rencontrer des ellipses du verbe principal ou des phrases nominales dans notre corpus (→ 4.2.3). Or, lorsque ces propositions sont en relation horizontale, nous observons des cas de parataxe entre propositions averbales ou entre une proposition verbale et une proposition averbale. Cela semble constituer une difficulté pour le parser ; SA-2 (fig. 31) en constitue un exemple.

Dans ce cas, tous les segments (séparés par des signes de ponctuation et dont les têtes sont *bouche*, *porte*, *tourbillon* et *refrain*) sont averbaux et typiques de notre recueil. En raison de leur caractère indépendant, nous pouvons considérer qu'il s'agit de propositions, dont le verbe est omis, en situation de parataxe. Les segments de *porte* et de *tourbillon* sont prédits comme une coordination à trois conjoints alors que *refrain* est considéré comme une apposition d'*eau*, et donc un dépendant du segment *tourbillon* introduisant le segment à un niveau syntaxique inférieur à celui des autres segments. L'analyse que nous en faisons est plutôt une parataxe à quatre éléments situés tous au

même niveau hiérarchique et dont le gouverneur[40] est *bouche*. La parataxe n'a donc pas été détectée, de même que la portée syntaxique d'un des segments, celui de *refrain*, a été sous-évaluée. Notons cependant que les prédictions du parser sont toutes des relations horizontales, ce qui limite l'erreur.

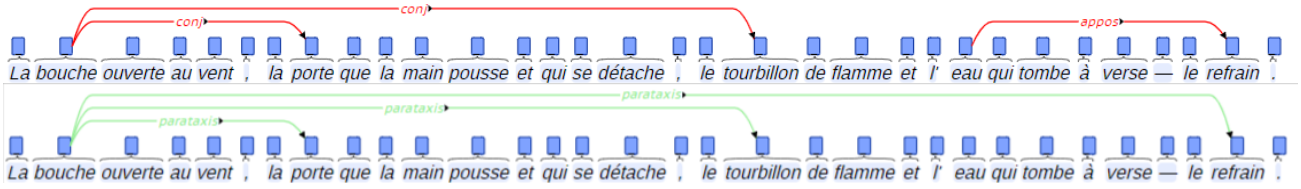


Fig. 31 — Prédiction et correction de SA-2

La situation apparaît également lorsque la première proposition est verbale, mais la seconde averbale ou inversement. Ce dernier cas est intéressant dans TN-3 (fig. 32) : la seconde proposition est étiquetée comme une proposition relative en l'absence de pronom relatif. Nous supposons que cela est dû au fait qu'une proposition dépendant d'un substantif est généralement une proposition relative de celui-ci.

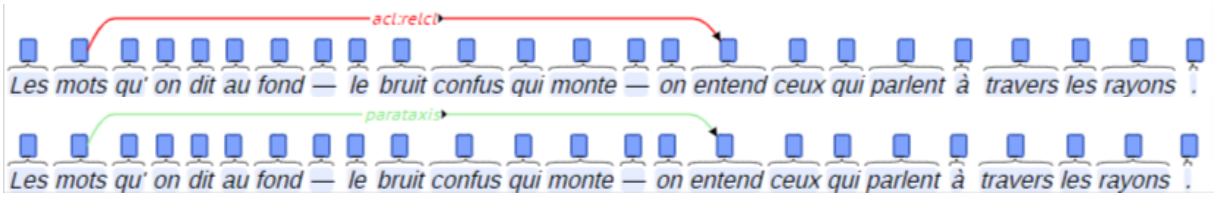


Fig. 32 — Prédiction et correction de TN-3

Nous rencontrons également 3 erreurs de confusion avec l'étiquette *advcl* (RI-2 [fig. 33], AP-1 et AP-7) ; le modèle prédit donc la proposition comme une proposition adverbiale. Cela est étonnant tout d'abord puisqu'il ne s'agit pas d'une relation horizontale. De plus, l'erreur est systématiquement incohérente, puisque le verbe de la proposition subordonnée adverbiale prédite est à un mode fini, mais la proposition n'est pas introduite par une conjonction de subordination, ce qui est incompatible : la proposition adverbiale est toujours introduite par une conjonction de coordination sauf lorsqu'il s'agit d'une proposition participe ou absolue, dont le mode du verbe est le participe (Grevisse et Goose, 2016 : 1598-1601). Nous n'avons aucune hypothèse concernant la raison de cette prédiction.

[40] Puisqu'il s'agit d'une relation horizontale, le gouverneur est un artefact et n'est présent que pour garantir que la structure reste un arbre.

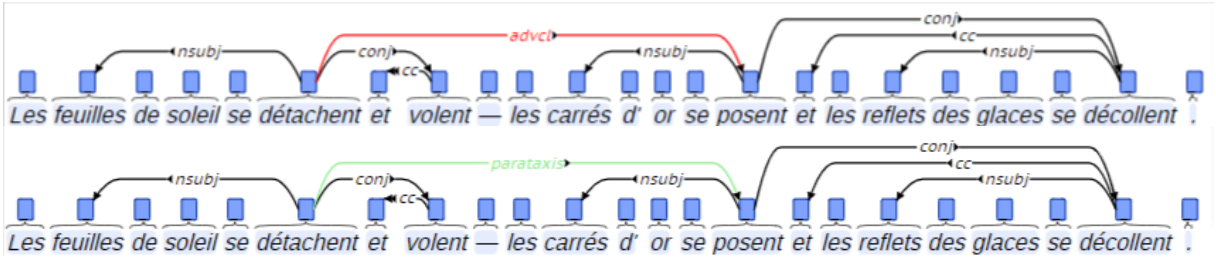


Fig. 33 — Prédiction et correction de RI-2

4.2.1.3 Apposition

Bien que la plupart des erreurs spécifiques à l'étiquette *appos* soient d'un type assez différent des précédentes, nous trouvons plus intéressant d'intégrer toutes les erreurs à cette sous-section.

L'analyse quantitative a permis de montrer une difficulté autour de l'étiquette *appos*, utilisée pour l'apposition, souvent en situation de faux négatif. Ceci a effectivement pu être confirmé lors de l'analyse manuelle : **26 erreurs** dans 19 phrases[41] sont liées à l'étiquette *appos*, et plus largement l'apposition, dont 23 faux négatifs. Il s'agit d'erreurs concernant en général l'étiquette et fréquemment incohérentes.

Dans le *Bon Usage*, l'apposition est définie comme « un élément nominal placé dans la dépendance d'un autre élément nominal et qui a avec celui-ci la relation qu'a un attribut avec son sujet, mais sans copule » (Grevisse et Goosse, 2016 : 462). À l'exception notable de l'apposition détachée, il s'agit généralement d'un groupe nominal placé à proximité d'un substantif et le déterminant à la manière d'un attribut. De cette façon, l'apposition n'est généralement pas précédée d'une préposition, à l'exception de trois types particuliers de constructions relevés par Grevisse et Goosse (2016 : 463-464) : les désignations objectives (*la ville* ***de Paris***), les désignations affectives (*son chef-d'œuvre* ***de robe***) et les tours littéraires où l'élément antéposé a une valeur métaphorique (*Je ne puis, Mégère libertine, / […] / Dans l'enfer* ***de ton lit*** *devenir Proserpine*). Nous pouvons donc considérer que les appositions en construction directe sont beaucoup plus fréquentes.

[41] PA-9, CS-5, VF-5, SA-5, R-3, BT-9, TE-5, TN-2, TN-3, NV-3, AP-3, GL-6, EF-3, TPB-1, TPB-4, TPB-5, TPB-10, TPB-14 et TPB-17.

Pourtant, le modèle semble confondre régulièrement l'apposition avec un complément déterminatif[42], qui se construit de façon indirecte. Le parser substitue ainsi l'étiquette *nmod* à l'étiquette *appos* dans 10 cas, comme illustré dans la fig. 34. Dans ce cas, rien ne nous permet d'expliquer la prédiction de complément déterminatif : aucun des substantifs dépendant de *Toi* n'est introduit par une préposition et il semble donc évident qu'il s'agisse d'appositions.

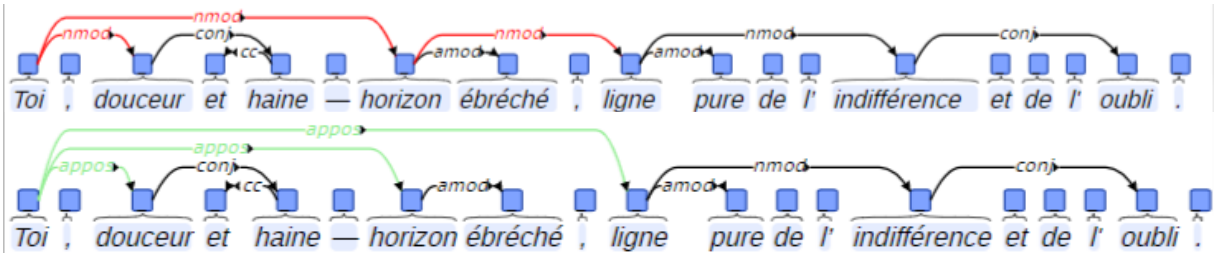


Fig. 34 — Prédiction et correction de TPB-14

L'hypothèse que nous pourrions formuler est que le parser n'a pas intégré correctement la notion de construction prépositionnelle et n'effectue pas de distinction claire entre un substantif duquel dépend une préposition et un substantif isolé. Cela peut être lié en partie au choix d'UD d'utiliser des mots lexicaux comme nœuds principaux des syntagmes plutôt que les éléments fonctionnels : lors de la prédiction de la dépendance du substantif, c'est alors seulement via les mécanismes d'attention que la préposition peut être repérée et considérée.

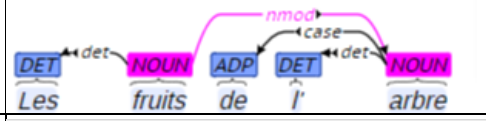
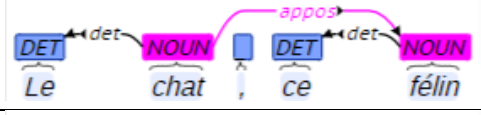
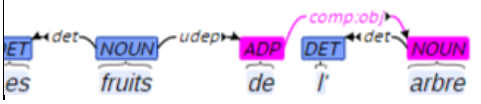
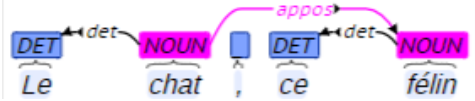
	Complément déterminatif	**Apposition**
UD		
SUD		

Fig. 35 — Complément déterminatif et apposition selon UD et SUD

Dans un format d'annotation privilégiant les têtes fonctionnelles, nous pouvons supposer que ce problème est moins fréquent, puisque la différence entre l'apposition et le complément déterminatif est beaucoup plus nette sur le plan syntaxique : le substantif complément déterminatif est ainsi dépendant d'une préposition là où le substantif apposition dépend d'un autre substantif. Il n'est donc pas possible pour le parser d'ignorer la préposition, comme l'illustre la figure 35 : dans SUD, la tête du

[42] *La maison* ***de mon voisin***. L'étiquette associée dans UD est *nmod* (*nominal modifier*).

complément déterminatif (*comp:obj*) est sa préposition. Les syntagmes prépositionnels sont différenciés par leur structure syntaxique même dans SUD, alors qu'il ne s'agit que d'une feuille *case* dépendant de la tête lexicale dans UD. Il est donc probable que ce type d'erreur ne puisse apparaître dans des prédictions basées sur SUD.

L'étiquette *conj* est la seconde à être le plus régulièrement substituée à *appos*, avec 5 occurrences. La fig. 36 illustre le cas de TPB-4. Dans ce cas, nous pouvons déduire qu'il s'agit d'une ambiguïté : est-ce que *l'esprit régulier* et *l'esprit commun* entretiennent une relation attributive avec *l'esprit de l'ordre*, en étant donc des appositions, ou est-ce qu'il s'agit d'une accumulation multipliant l'objet de *il y a* à l'aide d'une coordination ? Nos critères objectifs sont inefficaces dans ce cas, mais l'absence de conjonction de coordination entre les substantifs ainsi que la répétition précise de *l'esprit* indique plutôt une reformulation à nos yeux. Le parser, lui, a prédit la seconde option. Il est possible que la conjonction de coordination *Et* placée en début de phrase ait pu influencer faiblement la prédiction, mais cela est peu probable puisque le gouverneur de la conjonction est prédit correctement et que nous avons de nombreux exemples de conjonction de coordination orpheline n'ayant pas entraîné d'erreur dans une phrase. Nous soulignons cependant que, pour un modèle statistique, une différence minime peut exercer une influence suffisante pour que la prédiction s'en trouve altérée.

Il n'est pas étonnant que ces erreurs soient commises étant donné que l'opérateur humain ne peut pas toujours différencier l'apposition de la coordination avec certitude dans des cas spécifiques. Heureusement, comme nous l'avons souligné, les structures syntaxiques de l'apposition et de la coordination sont superposables dans UD et il ne s'agit donc que d'une erreur d'étiquette. Cela a néanmoins des répercussions sur la compréhension de la phrase.

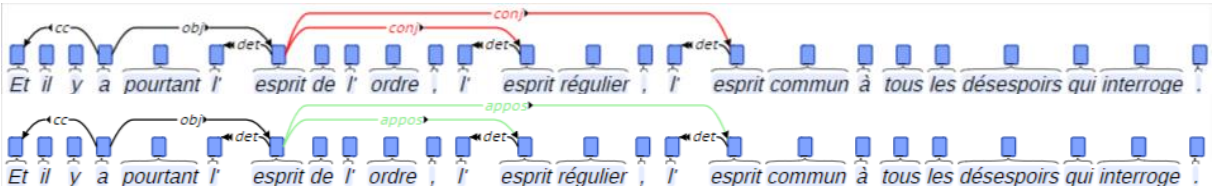


Fig. 36 — Prédiction et correction de TPB-4

D'autres étiquettes (comme *amod* [épithète], *obl:mod* [complément adverbial non essentiel] et *obj* [complément d'objet direct]) remplacent parfois *appos* également, mais il s'agit plutôt de cas isolés. Les cas d'erreur de sélection du gouverneur de l'apposition

s'éloignent des considérations de cette section et concernent plutôt la section intégrant les appositions détachées (→ 4.2.4).

4.2.1.4 Reprise syntaxique

Les erreurs concernant la reprise syntaxique sont peu nombreuses en raison de la faible fréquence (5) de ces structures. Elles sont généralement plutôt assimilables à des erreurs de déplacement ou détachement (→ 4.2.4) ou de projectivité (→ 4.2.14.1). Ces erreurs sont traitées dans ces sections.

4.2.2 Lexicalisation et tokenisation

Nous avons décidé de rassembler les erreurs dues au traitement analytique d'une unité lexicalisée et celles liées à une subdivision erronée d'un token par le tokenizer au motif qu'il serait pertinent dans les deux cas de rassembler les tokens concernés sous le même index dans l'arbre syntaxique. Cependant, nous les considérons distinctement dans les sous-sections suivantes puisqu'UD sépare nettement les lexicalisations (étiquette *fixed*) et les cas dans lesquels deux tokens n'en forment normalement qu'un (étiquette *goeswith* marquant une erreur de division à l'intérieur d'un mot dans un texte ou une erreur de tokenisation).

4.2.2.1 Difficultés liées aux unités lexicalisées

Comme le relèvent Sag *et al*. (2002 : 15), l'analyse des expressions à mots multiples (EMM, par exemple *à peine, l'on, goutte à goutte*…), constitue une difficulté du traitement automatique des langues, notamment à cause de leur diversité. En effet, certaines EMM n'ont plus aucune cohérence syntaxique interne et leur traitement analytique débouche généralement sur des incohérences. L'étiquette *fixed*, permettant d'analyser ces structures comme unitaires, est essentielle pour le bon fonctionnement de l'analyse, mais la difficulté vient souvent de la définition d'EMM : doit-on considérer uniquement les expressions figées comme *pendant que*, ajouter les expressions semi-figées du type *avoir l'air* ? Le premier n'admet aucun déplacement ou insertion et n'est plus analysable syntaxiquement, alors que le second admet l'insertion (*ne pas avoir l'air*, *avoir bien l'air*…) tout en s'intégrant à la catégorie des verbes copules dans son ensemble (→ 4.2.12.1).

Pour UD, il est important de limiter les expressions considérées comme des EMMs, mais aucun critère objectif n'est disponible pour aider l'annotation. Lorsqu'un ensemble de token est considéré comme EMM figée, le traitement est le suivant :

> Fixed MWEs [Multi-Word Expressions] are annotated in a flat structure, where all subsequent words in the expression are attached to the first one using the `fixed` label. The assumption is that these expressions do not have any internal syntactic structure (except from a historical perspective) and that the structural annotation is in principle arbitrary.[43]

Notons que les étiquettes *flat* et *compound* existent également pour l'analyse des EMMs, mais que nous n'en rencontrons pas dans notre corpus.

Durant la correction, nous avons relevé un ensemble d'EMMs qui auraient dû être considérées figées à nos yeux ; il s'agit des suivantes : *l'on* (C-3), *d'aplomb* (VF-10), *près de* (C-6, VV-9 et SA-6), *à peine* (PT-3, APN-1, R-3 et AP-8), *en bas* (VV-6), *pas même* (G-7), *d'autant* (CS-10), *sans que* (VF-1), *assez de* (VF-8 et VF-9), *pendant que* (SA-1), *pour que* (DC-5), *à part* (AP-1), *tête à tête* (GL-5) et *goutte à goutte* (EF-4). Le modèle a systématiquement tenté d'analyser ces unités, contrairement aux EMMs suivantes, que le modèle prédit correctement : *d'abord*, *un peu*, *de l'*, *peu à peu*, *à travers*, *plus de*, *tout à coup*, *tout à fait*, *à jamais*, *parce que* et *à demi*.

L'approche analytique de la prédiction de ces EMMs conduit très souvent à des situations absurdes, comme le cas de *l'on* (fig. 37). Le déterminant élidé *l'* n'a plus ici aucune valeur syntaxique ou sémantique et n'est qu'un vestige de la nature de substantif historique d'*on* (Grevisse et Goosse, 2016 : 1054). Il est donc curieux de voir cet élément *vide* prédit comme sujet nominal autonome, d'autant que le modèle prédit alors deux sujets qui ne sont pas coordonnés pour le verbe *ouvre*, ce qui est contraire au principe d'unicité du sujet.

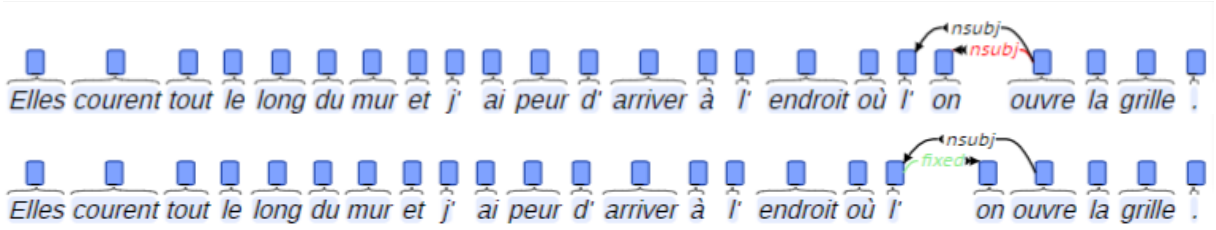


Fig. 37 — Prédiction et correction de C-2

[43] https://universaldependencies.org/u/dep/fixed.html

En plus des erreurs d'annotation interne, nous rencontrons de multiples cas dans lesquels le traitement analytique conduit à d'autres erreurs (fig. 38). Ainsi, les dépendances « *près* <advmod *vais* » et, dans une moindre mesure, « *près* obl:arg> *bois* » sont, pour nous, des conséquences de l'analycité de la prédiction de *près de*.

Fig. 38 — Prédiction et correction de C-6

Au vu du nombre d'EMMs qui n'ont pas été repérées, il n'est pas étonnant que **40 erreurs** réparties en 26 phrases soient concernées dans cette section. Aucune de ces erreurs n'est ambiguë, et nous constatons une répartition égale d'erreurs cohérentes et incohérentes (20 et 20).

Ce type d'erreur est bien évidemment lié à la particularité linguistique que sont les EMMs : il s'agit d'unités souvent tellement lexicalisées qu'il est devenu impossible de les décomposer syntaxiquement, mais elles sont pourtant écrites en plusieurs mots. Les seules solutions pour que le parser soit capable de réaliser des prédictions correctes sont l'intégration dans l'entraînement de tous les EMMs du français ou l'usage d'un dictionnaire externe pour le repérage de ces expressions. La difficulté réside alors dans la complexité de l'obtention d'un consensus des linguistes autour des EMMs.

4.2.2.2 Difficultés liées aux tokens subdivisés

Le traitement des tokens subdivisés (situations dans lesquelles un mot unique s'étale sur deux tokens, comme la division en deux tokens de *celui-là* en *celui* et *-là*) se rapproche de celui des EMMs : il est nécessaire de comprendre que plusieurs tokens forment un ensemble unitaire. La différence vient de l'origine de la difficulté : là où les EMMs sont le résultat de l'évolution normale de la langue au fil du temps, les tokens subdivisés apparaissent dans le cas d'erreurs typographiques (en particulier l'adjonction d'un blanc graphique dans un texte[44]) ou d'erreurs de tokenisation. Il s'agit dans notre cas d'un dysfonctionnement du tokenizer lié au trait d'union et à l'apostrophe. En effet,

[44] Ce type de token subdivisé peut également apparaître à l'oral, par exemple lorsque le locuteur marque une pause due à une difficulté d'élocution (*bien… venue*, dans un cas de déglutition par exemple) ou introduit un marqueur d'hésitation (*bien euh venue*).

dans le cas d'un tokenizer systématique, contrairement à la tokenisation par modèle statistique, un ensemble de règles et d'exceptions doivent être spécifiées pour que l'analyse lexicale soit réalisée correctement. Dans notre cas, il semblerait que les mots *amour-propre* (GL-6), *quelqu'un* (AR-1, G-6 et C-7), *peut-être* (VF-4 et C-9), *au-dessous* (TE-1) et *celui-là* (TN-2) manquent dans la liste des exceptions, puisqu'une division lexicale est opérée par le tokenizer après l'apostrophe (2 tokens, *quelqu'* et *un*) et autour du trait d'union (3 tokens, *amour*, - et *propre* ; à l'exception de *celui-là*, en 2 tokens [*celui* et *-là*], vraisemblablement à cause de la fréquence du clitique *-là*). Cela mène au même type d'erreurs que pour EMMs, parfois plus surprenantes puisque le trait d'union, lorsqu'il est constitué en token, reçoit généralement une fonction syntaxique (fig. 39).

Fig. 39 — Prédiction et correction de VF-4

Pour le traitement de ce type d'erreur, il est nécessaire d'habituer le parser à l'usage de *goeswith*, ce qui n'est pas le cas ici puisqu'une seule occurrence est présente dans Sequoia-Train. Cependant, puisque les subdivisions sont rarissimes à l'écrit, sauf erreur du tokenizer, un traitement de ces erreurs n'est peut-être pas pertinent au niveau du parser syntaxique. Ces considérations sont plus importantes dans le cadre d'un travail sur un corpus de langue orale.

Notons que **12 erreurs** sont concernées dans cette section.

4.2.3 Ellipse

À la lecture de ce recueil, il est impossible de ne pas être interpellé par des phrases comme VF-1 : l'auteur ne semble plus décrire une action, mais bien *déclarer* un objet ou une personne, en l'inscrivant dans une situation, mais sans lui assigner aucune action autre que celle d'*être*.

> (VF-1) Dans le coin le plus sûr, et sans qu'il y paraisse, la main du songe creux,
> au coin de l'éventail.

Ce style, que nous pourrions qualifier de *description ontologique*, nécessite l'absence d'un prédicat dans la phrase, et donc l'absence d'un verbe principal. Cependant, contrairement aux phrases nominales, l'auteur ne s'astreint pas à un

syntagme nominal seul, mais lui ajoute des compléments de phrase, adverbiaux ; cela est syntaxiquement impossible, sauf à supposer l'ellipse du verbe prédicatif, gouverneur des compléments adverbiaux. Du point de vue syntaxique, le verbe principal est donc régulièrement omis.

Évidemment, ce phénomène n'est pas anodin : pour analyser syntaxiquement une phrase dont le verbe est effacé, il est nécessaire de comprendre, par le contexte, l'absence de ce verbe et le supposer durant l'analyse ; il s'agit d'une opération particulièrement complexe et abstraite pour une machine. Il n'est donc pas étonnant, au vu de la quantité d'ellipses, que le présent type d'erreur soit un des plus fréquents dans le corpus.

Dans UD[45], il n'est pas possible de déclarer un token fantôme qui permettrait de symboliser les éléments dont l'ellipse est faite. Des mécanismes alternatifs sont donc nécessaires, ce qui est réalisé par la promotion d'un dépendant du nœud effacé, selon une hiérarchie imposée pour la cohérence. Les autres dépendants du nœud effacé deviennent alors dépendants du token promu. Afin de conserver la cohérence et éviter des ambiguïtés lorsqu'un verbe est élidé et qu'un des actants — généralement le sujet — est promu à sa place, les autres actants ainsi que les circonstants dépendant de ce verbe reçoivent l'étiquette *orphan*. Les mots fonctionnels, cependant, conservent leur étiquette originelle, mais doivent dépendre également du nœud promu (voir fig. 40). Les erreurs de cette section s'organisent donc autour d'un type, les erreurs liées à une ellipse (en particulier l'ellipse d'un verbe), et un sous-type, celui des erreurs de l'étiquette *orphan*.

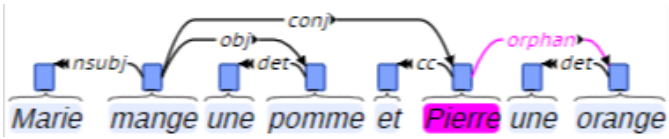


Fig. 40 — Promotion du sujet et conséquence sur l'objet

Nous rencontrons 64 erreurs liées à une ellipse, réparties en 25 phrases[46]. Parmi celles-ci, 46 erreurs sont liées à l'étiquette *orphan*, qui n'a pas été prédite une seule fois par le parser, ce qui n'est pas étonnant en raison de sa faible fréquence dans Sequoia-

[45] https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis

[46] C-2, AR-1, AR-7, PT-1, PT-2, PT-3, FP-1, FP-3, G-7, G-8, CS-1, CS-2, VF-1, VF-2, VF-5, VF-9, SA-2, SA-4, SA-6, DC-3, BT-2, BT-8, BT-11, TE-1, AA-4, HE-1, HE-2, HE-6, HE-7, AP-8, AP-13, GL-1, GL-2, GL-3 et TPB-15.

Train (31 occurrences) et les difficultés qui y sont associées. La densité de ces erreurs est assez élevée puisqu'il s'agit, en moyenne, de 2,56 erreurs par phrase reprise dans cette section.

Le tableau ci-dessous montre la répartition des 64 erreurs en fonction de leur type et de leur catégorie de typologie grossière.

Type d'erreur	**HEAD**	**DEPREL**	**BOTH**
Nombre d'erreurs	2	29	33

Tab. 7 — Répartition des erreurs d'ellipse

Nous observons une faible quantité de prédictions correctes de l'étiquette alors que la prédiction du gouverneur ne l'est pas (HEAD). Cela n'est pas particulièrement étonnant au vu du nombre de faux négatifs de l'étiquette *orphan* dans ce cas. La seule possibilité d'apparition de cette erreur, sauf hasard, est une situation dans laquelle le token à prédire est un mot fonctionnel et le nœud promu est erroné. Or, puisque le nœud à promouvoir est en général le plus proche des mots fonctionnels, il est assez naturel pour le modèle de le sélectionner comme gouverneur de ceux-ci.

C-2 (fig. 41) présente une erreur ambiguë liée à une ellipse, ce qui change la dynamique syntaxique. Nous remarquons dès la première lecture la présence d'un chiasme : dans la première section (en cyan), *têtes* suit *arbres* en étant son attribut alors que dans la seconde (en magenta) c'est *arbres* qui suit *têtes*, mais en l'absence de la copule. Évidemment, grâce à la similarité de la construction, nous supposons aisément l'ellipse de la copule (*Les arbres sont des têtes, ou les têtes* [*sont*] *des arbres*). De plus, coordonner l'attribut n'a que peu de sens : dire que l'arbre est soit une tête, soit une tête d'arbre est peu cohérent.

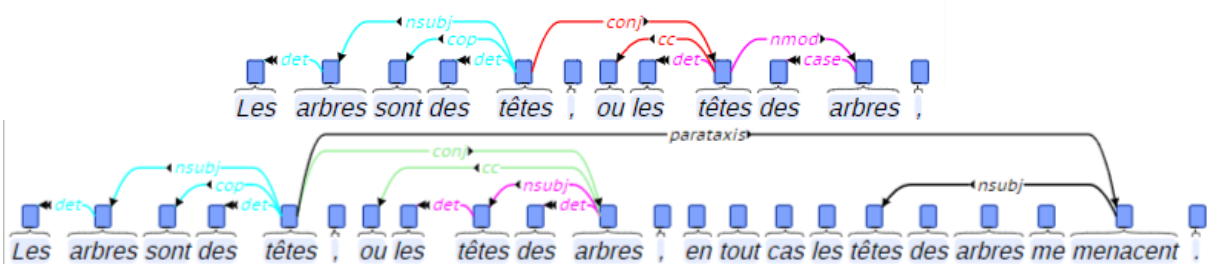


Fig. 41 — Prédiction et correction de C-2

La solution proposée par le parser, c'est-à-dire une structure se projetant vers la droite avec l'allongement du second membre de la coordination à l'aide d'un complément déterminatif (*des arbres*), est syntaxiquement correcte et évidente lorsque

nous ne supposons pas l'ellipse, mais elle ne semble pas convenir sur le plan sémantique et stylistique. Dans ce cas, il faut donc comprendre que la copule est omise et répéter la première structure en sélectionnant le second substantif comme attribut. Plusieurs erreurs sont donc liées : la coordination (et donc le gouverneur de la conjonction de coordination *ou*) est erronée parce que le verbe est élidé ; le premier substantif disponible est donc sélectionné, et la structure est forcée de se projeter vers la droite, contrairement à la construction attributive qui se projette plutôt vers la gauche avec le sujet dépendant de l'attribut.

Notons également que C-2 présente une erreur de segmentation (→ 4.2.14.2).

Dès l'évaluation automatique, nous avons pu constater que l'étiquette *nmod* est souvent en situation de faux positif, c'est-à-dire qu'elle est prédite, mais est erronée. Une des raisons de cette surreprésentation de *nmod* se trouve dans les erreurs d'ellipse : parmi les erreurs sélectionnées dans cette section, 24 tokens étaient prédits *nmod* mais l'étiquette n'est présente aucune fois dans la correction ; il y a donc 24 faux positifs rien que pour les erreurs d'ellipse. AA-7 (fig. 42) illustre parfaitement ce phénomène, avec 3 prédictions erronées de *nmod*, plutôt cohérentes, voire ambiguës, à la place de l'étiquette *orphan*.

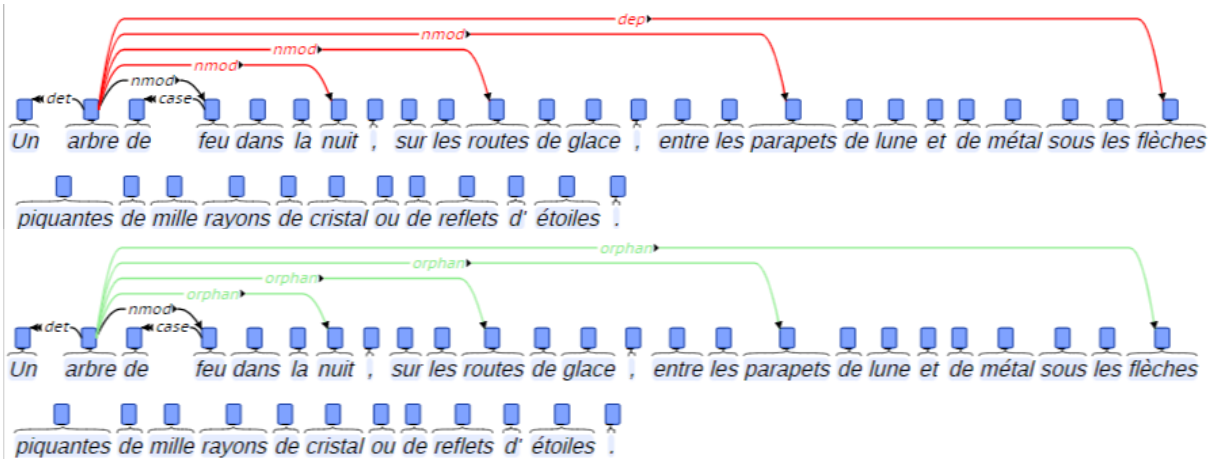


Fig. 42 — Prédiction et correction d'AA-7

Le comportement du modèle est justifié dès lors que l'on considère qu'il n'est pas capable de supposer l'ellipse : les compléments adverbiaux à tête nominale (*dans la nuit*, *sur les routes de glace*, *entre les parapets de lune et de métal* et *sous les flèches piquantes de mille rayons de cristal ou de reflets d'étoiles*) ont en effet la forme attendue d'un complément déterminatif, puisqu'ils sont constitués d'un syntagme

nominal introduit par une préposition. Entre *nmod* et *obl:mod* (complément adverbial non essentiel), il reste une différence de nature du gouverneur : dans le cas du complément déterminatif, le gouverneur est nominal ; pour le complément adverbial, il est verbal. Puisque le verbe est omis et que le nœud promu est nominal, les compléments qui devaient dépendre d'un verbe se mettent à dépendre d'un substantif, ce qui supprime cette distinction syntaxique majeure entre *nmod* et *obl:mod*. Il est curieux que le quatrième complément, *sous les flèches […] étoiles* reçoive l'étiquette *dep* dans la prédiction malgré la sélection correcte du gouverneur et l'étiquette des trois compléments précédents. La longueur de la dépendance ou l'absence de ponctuation avant ce complément, contrairement à *sur les routes de glace* et *entre les parapets de lune et de métal*, peuvent être des facteurs explicatifs.

BT-11 (fig. 43) montre une erreur de sélection de la racine probablement liée à une ellipse. En l'absence d'un verbe à un mode fini, le modèle prédit généralement un substantif comme racine. Cependant, dans ce cas, le substantif *fond* fait partie d'un syntagme prépositionnel, et seul le substantif *renard* est un candidat plausible comme racine, d'autant qu'il s'agirait de la tête promue en cas d'ellipse. Nous pouvons supposer que l'erreur est due à l'absence de verbe à un mode fini combinée à une racine apparaissant assez loin dans la linéarité de la phrase. Nous développons plus ce type d'erreur dans la section des erreurs de racine (→ 4.2.5). Notons que la dynamique de la phrase est bouleversée : d'une phrase plutôt vers l'arrière, puisque la racine est tardive, le modèle prédit une structure syntaxique en majorité vers l'avant.

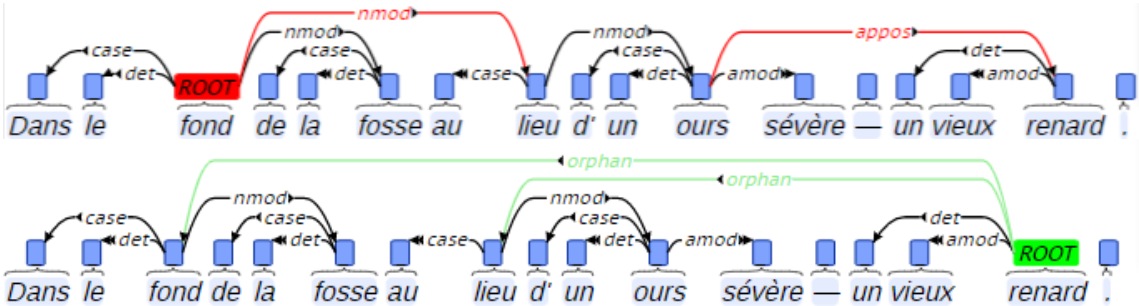


Fig. 43 — Prédiction et correction de BT-11

Nous avons pu constater trois grandes difficultés de l'ellipse, en particulier l'ellipse du verbe : la sélection de la racine, le choix du token à promouvoir et les ressemblances syntaxiques entre certains compléments. Cela conduit le modèle à se tromper régulièrement lorsque le verbe principal est manquant, d'autant que la détection d'une ellipse est conditionnée par un processus abstrait et la prise en compte du contexte

entier. Il s'agit d'un phénomène linguistique face auquel le modèle, à moins peut-être d'un entraînement ou de modèles spécifiques, est démuni en raison de l'abstraction nécessaire à la prise en compte d'un élément omis.

4.2.4 Adjonction ou déplacement d'un complément

Nous rassemblons dans cette section toutes les erreurs qui sont vraisemblablement liées au déplacement d'un complément ou à l'adjonction d'un autre élément entre le syntagme et le dépendant, ce qui complexifie sa prédiction en allongeant les dépendances. Cette section concerne 38 erreurs réparties dans 25 phrases[47], ce qui en fait un des types les plus répandus dans notre corpus.

Nous rencontrons en particulier le phénomène de déplacement dans trois cas : l'épithète détachée (9 erreurs, dont 7 adjectifs [étiquette *amod*] et 2 propositions adjectivales [*acl*]), l'apposition détachée (12 erreurs [*appos*]) et la réduplication du sujet, ou reprise syntaxique, en début ou fin de phrase (5 erreurs [*dislocated*]). Le tableau 8 reprend le nombre d'erreurs par type.

Type d'erreur	**HEAD**	**DEPREL**	**BOTH**
Nombre d'erreurs	6	5	23
Typologie grossière	**Incohérent**	**Cohérent**	**Ambigu**
Nombre d'erreurs	23	9	2

Tab. 8 — Répartition des erreurs d'adjonction ou déplacement

Nous constatons que la plupart des erreurs concernent au moins la sélection du gouverneur, puisque seules 5 erreurs ne se situent qu'au niveau de l'étiquette (DEPREL). Celles-ci sont d'ailleurs toutes incohérentes. Il n'est pas étonnant que la sélection du gouverneur soit malaisée : nous avons posé comme caractéristique de cette section un déplacement ou une insertion, ce qui implique souvent un allongement des dépendances et le gouverneur est donc souvent remplacé par un token plus proche dans la phrase. Cela semble se confirmer lors de l'observation de la différence de moyenne de longueur des dépendances des tokens concernés par une erreur du présent type, comme nous pouvons le voir dans le tableau ci-dessous.

[47] PA-1 (1), CS-10 (1), VF-10 (1), SA-5 (1), BT-9 (1), TE-1 (1), TE-5 (2), TE-6 (2), LNP-3 (1), LPN-4 (2), LPN-7 (1), TN-2 (1), AA-5 (1), NV-1 (1), NV-3 (1), NV-6 (1), NV-8 (2), NV-14 (1), AP-2 (1), AP-3 (2), AP-4 (1), GL-6 (2), EF-1 (1), TPB-1 (4), TPB-5 (1) et TN-5 (4).

	Avant la correction	Après la correction
MLD des erreurs de cette section	5,5151	11,8484

Tab. 9 — MLD des erreurs d'adjonction ou de déplacement

La différence est spectaculaire puisque nous constatons qu'en moyenne le parser attribue le dépendant à un token deux fois plus proche de lui que ne l'est le gouverneur correct.

Nous pouvons également voir que plus des deux tiers des erreurs sont incohérentes, et seules 2 sont ambiguës. Le déplacement et l'insertion semblent donc être des difficultés particulièrement importantes pour le parser. De même, en regardant les étiquettes des prédictions erronées, il est très difficile d'y déceler une systématicité et le comportement a parfois l'air aléatoire. Nous nous contentons donc d'illustrer ces erreurs par des exemples et d'en décrire les particularités syntaxiques.

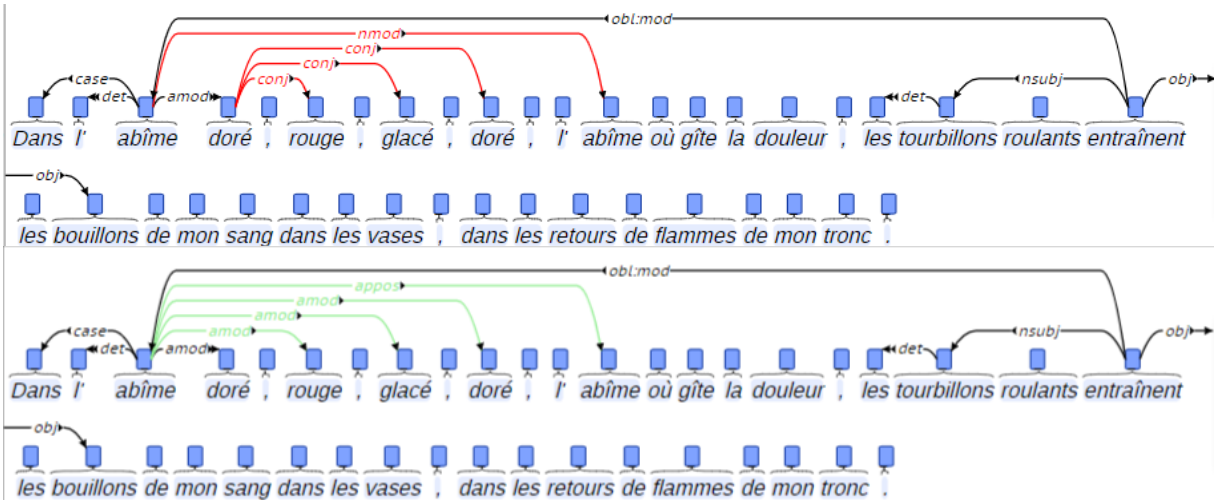


Fig. 44 — Prédiction et correction de TPB-1

TPB-1 (fig. 44) est le premier exemple sur lequel nous aimerions attirer l'attention du lecteur. Nous pouvons y constater une erreur d'apposition détachée (*« *abîme* nmod> *abîme* ») ainsi que trois erreurs se rapprochant de la situation d'une épithète détachée, mais aussi des piles paradigmatiques : les adjectifs *rouge*, *glacé* et *doré* sont séparés du substantif duquel ils dépendent par des virgules. Concernant l'apposition détachée, le comportement est similaire à celui déjà décrit : l'étiquette *nmod* (complément déterminatif) est substituée à l'étiquette *appos* (apposition). Pourtant, le fait que le substantif *abîme* soit répété renforce encore la probabilité que le syntagme soit une apposition puisqu'il est courant, dans ce type de construction, de surqualifier un substantif en le répétant. Ensuite, les erreurs concernant les adjectifs sont toutes

cohérentes : il est possible qu'un substantif soit déterminé par des épithètes coordonnées. Cependant, puisque la coordination est implicite et que la fonction épithète peut être multipliée, nous considérons qu'il s'agit plutôt d'une pile paradigmatique. Le (faible) détachement matérialisé par l'usage de la virgule a pu tromper le parser à ce sujet en le poussant à coordonner les épithètes.

VF-10 (fig. 45) illustre un cas clair d'épithète détachée (***Ivre**, en rentrant chez lui, Pierre est tombé*). En effet, *d'aplomb* s'est lexicalisé et se comporte désormais comme un adjectif : pour le prouver, il suffit d'utiliser la commutation et remplacer *d'aplomb* par un membre du paradigme des adjectifs, par exemple *vertical*. Puisque *Et, toujours vertical, le masque est impassible* est tout à fait cohérent, nous pouvons conclure que *d'aplomb* occupe la fonction d'épithète, détachée dans ce cas puisqu'il est placé avant l'actualisateur (*le*) du substantif qu'il détermine (*masque*). Le modèle, lui, a considéré *toujours d'aplomb* comme un complément adverbial non essentiel, dépendant de la racine de la phrase, c'est-à-dire l'attribut du sujet dans une construction à copule (→ 4.2.12.1). L'erreur peut s'expliquer par le déplacement en tête de phrase, ce qui est souvent une place de complément adverbial, mais aussi le fait que le modèle n'ait pas repéré la lexicalisation. Il était alors beaucoup plus malaisé de considérer qu'*aplomb* détermine *masque* puisque la seule possibilité, en construction indirecte, est le complément déterminatif et il est très rare que ce complément soit détaché et antéposé. Remarquons qu'ici la dépendance est plus courte après la correction et le modèle ne s'est donc pas contenté du token disponible le plus proche.

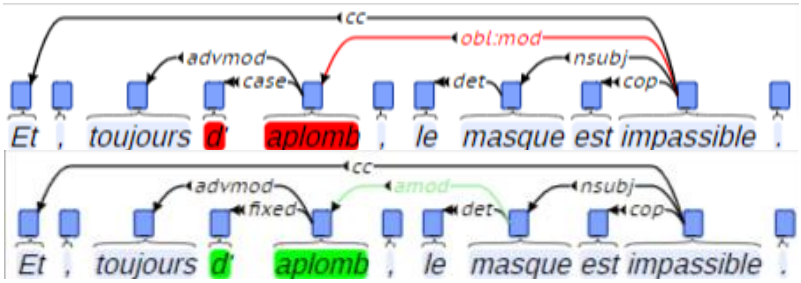


Fig. 45 — Prédiction et correction de VF-10

Enfin, LPN-4 (fig. 46) illustre une série d'erreurs liées à l'adjonction d'un complément dans une structure syntaxique. Le complément adverbial *là où on ne mesure plus la profondeur* (en cyan dans la correction) est intercalé avant le verbe auquel il se rapporte et son contexte (*et il y a des rencontres imprévues…*, en magenta),

lui-même dans une situation de coordination à trois éléments avec la racine de la phrase, *détache*, et une proposition tierce (les *dessous humides s'unissent*, en rouge).

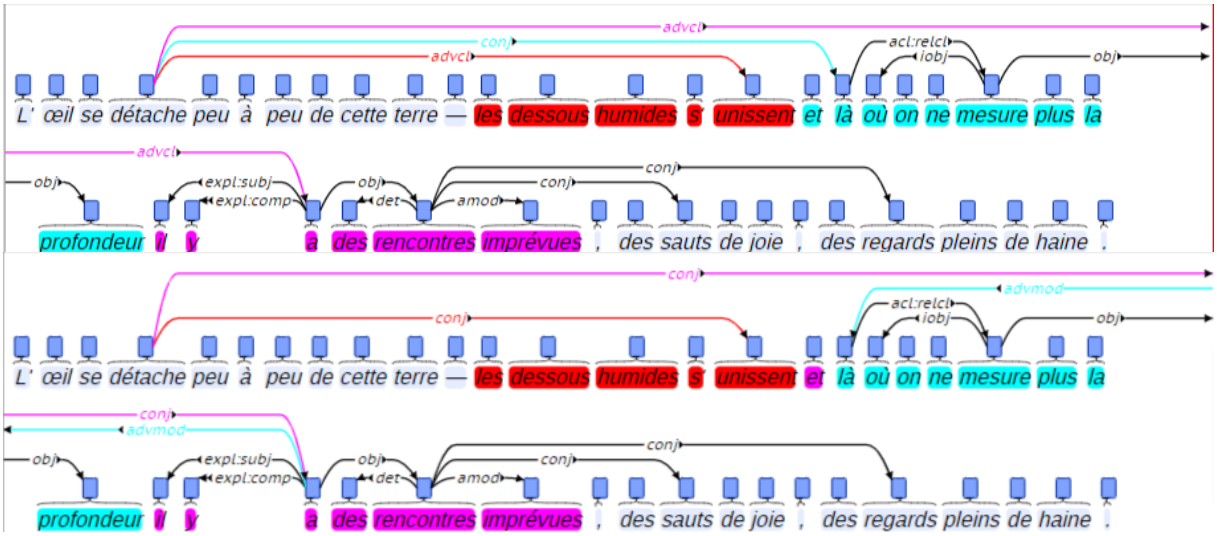


Fig. 46 — Prédiction et correction de LPN-4

Nous pouvons constater que cet élément a constitué une difficulté pour le modèle puisque c'est lui qui est considéré comme élément coordonné grâce à la conjonction *et*, alors que la tête *là*, qui a un statut adverbial, ne peut être coordonnée avec le verbe fini *détache* sans supposer l'ellipse du verbe. À la suite de cela, la proposition *les dessous humides s'unissent* n'est plus considérée comme membre de la coordination et est prédite proposition adverbiale, tout comme *il y a des rencontres imprévues*, qui perd la conjonction de coordination *et* puisqu'elle est attribuée ailleurs. Dans la correction, nous observons effectivement la présence d'une proposition intermédiaire après la conjonction, mais celle-ci a un statut adverbial et se rapporte à la proposition qui la suit.

Le modèle a lié ici la conjonction de coordination avec le verbe le plus proche, ce qui est erroné lorsque le premier membre de la coordination n'appartient pas au même niveau propositionnel que le second : *là où on ne mesure plus la profondeur* ne peut être proposition principale coordonnée à *L'œil se détache peu à peu de cette terre*, contrairement à *il y a des rencontres imprévues*… C'est donc cette dernière proposition qui est le dernier conjoint, malgré la distance avec la conjonction *et*. Nous émettons l'hypothèse que la prédiction erronée de la section *les dessous humides s'unissent* est soit un effet de bord de la prédiction de la coordination, soit une conséquence de la présence du tiret cadratin.

4.2.5 Sélection de la racine de la phrase

Le nœud racine (*root*) occupe une place particulière : il est nécessaire et unique, c'est-à-dire qu'il y a toujours une et une seule racine dans une phrase. Ce nœud est très important, puisque c'est autour de celui-ci que se construit tout l'arbre syntaxique. Lorsque la racine est un verbe par exemple, c'est de lui que dépendent tous les actants (sujet, éventuels compléments d'objet direct et d'objet indirect), les circonstants (compléments adverbiaux), mais aussi des éléments comme l'auxiliaire ainsi que la tête d'éventuelles propositions coordonnées. Nous comprenons donc que la prédiction incorrecte de la racine de la phrase entraîne des conséquences importantes. En effet, une prédiction de racine erronée entraîne *a minima* deux erreurs BOTH (le gouverneur et l'étiquette de la racine prédite et de la racine réelle), et souvent des erreurs supplémentaires puisque la structure la plus fondamentale de la phrase n'est pas reconnue. Puisque la prédiction de 12 racines[48], sans compter les racines dues aux erreurs de segmentation, est erronée, nous relevons un total d'au moins **35 erreurs** liées à une erreur de sélection de racine, c'est-à-dire 24 erreurs directes et au moins 11 erreurs conséquentielles. Celles-ci peuvent être illustrées par PA-1 (fig. 47) et NV-3 (fig. 48).

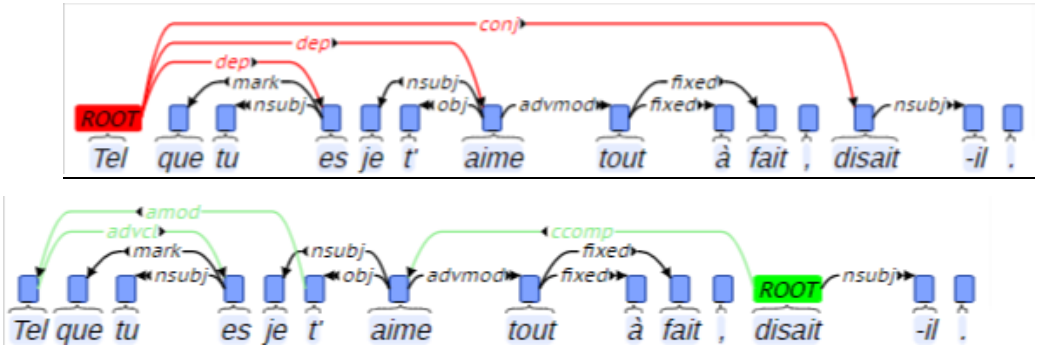


Fig. 47 — Prédiction et correction de PA-1

Dans le cas de PA-1 (fig. 47), la sélection de *Tel* comme racine est particulièrement étonnante et très probablement due à de nombreux facteurs (inversion sujet-verbe du verbe principal, verbe principal en fin de phrase, difficultés de projectivité [→ 4.2.14.1]…). Cela conduit à une analyse vers l'avant, avec une dynamique d'éléments plutôt postposés, alors que l'arbre syntaxique correct est plutôt vers l'arrière et contient des éléments antéposés. De plus, la prédiction montre un bouquet et trois sous-arbres

[48] PA-1, PT-3, FP-1, CS-1, VF-9, APN-6, BT-2, BT-11, TE-1, LPN-3, TN-2 et NV-3.

disjoints dépendant de la racine (*que tu es, je t'aime tout à fait* et *disait-il*) alors que l'analyse correcte présente un seul sous-arbre (*Tel que tu es je t'aime tout à fait*).

L'erreur de racine de NV-3 (fig. 48) est étonnante puisqu'un verbe à un mode fini et non accompagné de conjonctions (*semblait*) est présent et semble être le candidat idéal ; c'est pourtant un verbe à un mode non fini (*Écrit*) qui est choisi. Il est d'ailleurs étonnant que la proposition *il semblait vouloir franchir la haie […]* soit prédite comme une proposition subordonnée adverbiale alors qu'aucune marque ne peut l'indiquer.

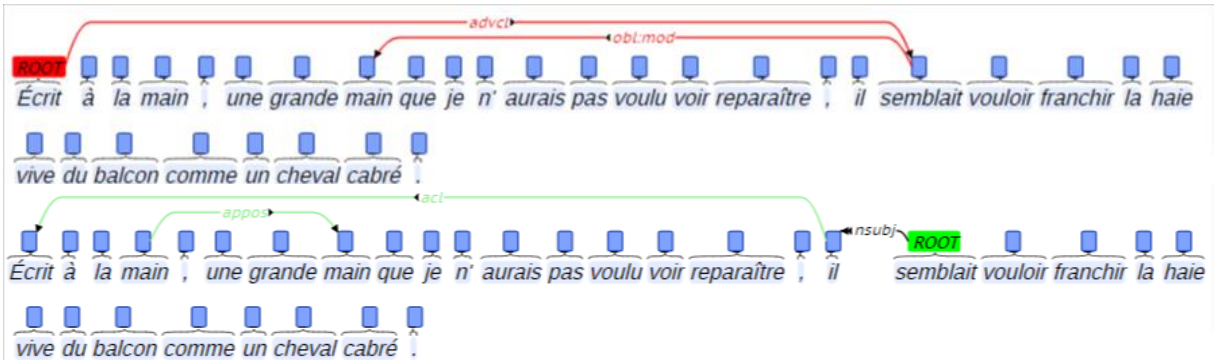


Fig. 48 — Prédiction et correction de NV-3

Notons que dans les deux exemples précédents, la racine prédite apparaît bien plus tôt dans la linéarité de la phrase que la racine réelle, ce qui indiquerait que le modèle a plus de difficultés à trouver la racine lorsqu'elle est précédée d'un plus grand nombre d'éléments. Cette tendance semble se vérifier dans l'ensemble des phrases concernées dans cette section : parmi les 12 exemples, dans 11 cas la racine prédite précède la racine réelle dans la phrase. Seule la racine prédite de TN-2 (*-là*) suit la racine réelle (*Celui*), mais cela est dû à une erreur du tokenizer : la racine est en réalité *Celui-là*, mais une division a été effectuée ; par convention, c'est le premier des deux éléments qui reçoit la relation syntaxique réelle et les suivants la relation *goeswith*.

> (TN-2) Celui-là — le corps et la tête et l'âme dans l'espace — tout ce qui dure encore — et traîne sur le soir.

La différence est particulièrement marquée au regard de la position moyenne des racines (moyenne des ID des tokens *root*) :

Racines	Racines erronées prédites	Racines erronées corrigées	Racines correctement prédites (corpus entier)
Position moyenne	4,5833	18,8333	4,7681

Tab. 10 — Position moyenne des racines dans le corpus

Il ne s'agit pas d'une règle systématique puisque, par exemple, la racine d'EF-3 qui apparaît à la quarante-septième position est prédite correctement, mais il s'agit bien là d'une tendance nette.

> (EF-3) Par toutes les épines qui sillonnent les routes du cœur et de la pensée, chemins coupés de meurtrissures, de rives d'eau, de colliers de larmes et de signes, tracés par la haine et le ressentiment des bêtes, je ne me **reconnais**$_{\text{ROOT}}$ pas dans ces pages au miroir méfiant de la source.

Selon nous, le fait que la racine apparaisse loin dans la phrase est un facteur important, mais insuffisant dans la plupart des cas. Nous remarquons en effet que la plupart des phrases dont la racine prédite est erronée comportent des particularités qui ne leur sont pas exclusives, mais qui s'additionnent à la racine lointaine : une ellipse, une antéposition particulière ou une incise.

Dans le cas de PT-3, ainsi que 6 autres phrases (FP-1, CS-1, VF-9, BT-2, BT-11 et TN-2), la phrase présente une ellipse du verbe marquée par la présence de compléments de phrase, mais pas de verbe qui les gouvernent.

> (PT-3) Sur les **revers**$_{\text{PRÉDIT}}$ à peine revenus, aux carrefours émus par le passage, les **battements**$_{\text{CORRECT}}$ de pieds au sol interrompus, le souffle blanc dans les buissons qui se dispersent.

Pour APN-6, la racine prédite est la tête d'une proposition incidente jouant un rôle de parenthèse et encadrée par des tirets cadratin.

> (APN-6) Au climat desséchant de cette pente, à la place de ce poids humide du bas-fond, malgré le mouvement des peupliers inquiets qui se le disent — il ne **peut**$_{\text{PRÉDIT}}$ y avoir de place pour le cœur gonflé qui se repent — on **tranche**$_{\text{CORRECT}}$ la noirceur de l'esprit qui s'envole.

Enfin, dans le cas de TE-1, NV-3 et LPN-3, une épithète détachée du sujet ou une répétition de celui-ci — dans une situation de reprise syntaxique — est placée en tête de phrase.

4.2.6 Essentialité des compléments de phrase

Pour **27 erreurs** réparties en 23 phrases[49], le parser a confondu un complément essentiel et un complément non essentiel. Selon Grevisse et Goosse (2016 : 346), les compléments sont essentiels **1.** « quand leur construction (présence ou non d'une

[49] PA-6, C-3, AR-8, PT-1, PT-2, PT-5, PT-6, G-2, CS-2, CS-5, CS-10, CS-11, SA-5, SA-6, DC-5, R-1, BT-5, AA-2, AA-7, NV-1, NV-11, AP-4, AP-6, EF-3, EF-4 et TPB-10.

préposition, choix de la préposition) dépend du verbe lui-même » ; **2.** « quand le verbe ne peut constituer sans eux le prédicat ». De cette façon, *à Paris* est essentiel dans *Pierre va à Paris*, car le verbe impose la présence d'une préposition (**Pierre va Paris* mais *Pierre va vers Paris*) et ne peut être présent seul (**Pierre va*). Au contraire, il est non essentiel dans *Pierre mange à Paris*, puisque *Pierre mange* et *À Paris, Pierre mange* sont cohérents.

L'opposition entre les étiquettes de compléments prépositionnels dépendant d'un verbe *obl:mod* et *obl:arg* est justement fondée sur l'essentialité : *obl:mod* est l'étiquette qualifiant la relation au verbe d'un complément non essentiel alors qu'*obl:arg*, celle d'un complément essentiel. L'analyse quantitative a fait apparaître une grande quantité de faux positifs de la relation *obl:arg*, et cela s'explique par cette section : de nombreux compléments sont prédits comme essentiels par le parser alors qu'ils ne le sont pas. En effet, dans 20 cas, le parser a prédit l'étiquette *obl:arg* pour un circonstant prépositionnel non essentiel (étiquette *obl:mod*), comme illustré dans la figure 49. Le cas contraire n'apparaît que 2 fois.

Dans l'exemple ci-dessous, nous pouvons en effet constater que le complément *contre les murs qui frissonnent* n'est pas essentiel : *La voiture habituelle roule, Contre les murs, la voiture habituelle roule* et *La voiture habituelle roule sur les murs* sont des phrases tout à fait correctes.

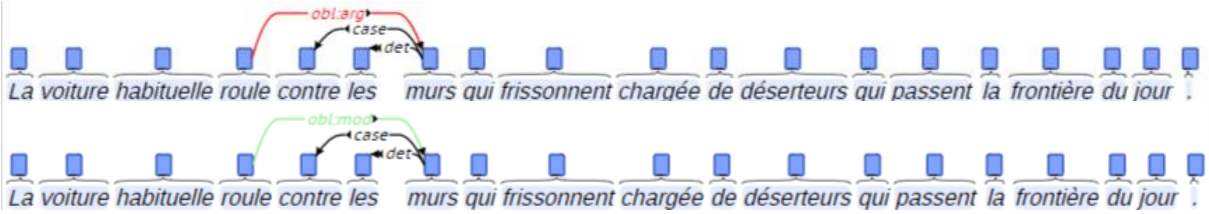


Fig. 49 — Prédiction et correction d'AP-4

La quasi-totalité (24 sur 27) des erreurs sont uniquement des erreurs DEPREL, mais sont néanmoins importantes : puisqu'*obl:arg* est également l'étiquette assignée au complément d'objet indirect lorsqu'il n'est pas pronominal[50], nous pouvons soupçonner une confusion entre complément adverbial essentiel, complément adverbial non essentiel et complément d'objet indirect de la part du parser. Néanmoins, le caractère essentiel d'un complément est parfois difficile à établir et il est donc compréhensible

[50] Pour les corpus du français, UD différencie en effet *Pierre **lui** parle*, annoté *iobj*, et *Pierre parle **à Marie***, *obl:arg*.

que le parser éprouve des difficultés à cet égard. Il est étonnant cependant que la prédiction *par défaut* soit celle d'un complément essentiel pour les compléments adverbiaux puisque le complément adverbial non essentiel est très fréquent. Nous soupçonnons que ceci soit dû au fait que l'étiquette du complément d'objet indirect soit également *obl:arg*, augmentant la fréquence de cette étiquette. Malgré cela, *obl:mod* reste plus fréquente dans le corpus d'entraînement avec 1670 occurrences, contre 1092 pour *obl:arg*.

Puisque la liste des verbes admettant un complément d'objet indirect ou un complément adverbial essentiel est limitée, nous pourrions nous attendre à plus de facilités pour la prédiction. La proximité des trois structures — qui sont toutes des syntagmes prépositionnels dépendant d'un verbe — est probablement un facteur de confusion.

4.2.7 Compléments prépositionnels d'un adjectif ou d'un adverbe

Lors de la correction, nous avons pu remarquer que la plupart des compléments prépositionnels subordonnés à un adjectif (Grevisse, 2016 : 493) sont problématiques pour le parser ; cela résulte en **14 erreurs** réparties en 11 phrases[51].

La particularité des compléments de l'adjectif pour UD est qu'aucune étiquette n'est prévue pour spécifier la relation hormis *advmod* permettant de traiter les adverbes (***très*** *gentil*). Dans le cas de compléments prépositionnels comme *pleines* ***de vase*** (AP-8), il serait absurde de considérer que le modifieur *de vase* est résumable à un adverbe seul comme le nécessite *advmod* : il est impossible de commuter ce syntagme prépositionnel avec un élément du paradigme des adverbes sans compromettre la phrase avec un résultat du type **pleines* ***très*** ou **pleines* ***toujours***. Les compléments prépositionnels d'adjectifs ne peuvent alors être étiquetés qu'au moyen de la relation sous-spécifiée *dep*.

> (AP-8) Il fuit le silence hébété, à peine dégagé des rayons lumineux des roues de la tempête — les ornières de sa destinée pleines de vase.

Nous relevons trois types d'erreurs concernés par cette section : 7 erreurs DEPREL substituant l'étiquette *obl:arg* à *dep*, 4 erreurs d'ambiguïté présentant de faux positifs de

[51] PT-4, VV-9, APN-5, APN-7, LPN-4, AA-4, HE-6, AP-3, AP-7, AP-8 et TPB-5.

nmod et *obl:mod*, ainsi qu'un cas de figement (*près de*) pris pour un adverbe avec complément entraînant 3 erreurs.

PT-4 (fig. 50) illustre une erreur substituant *obl:arg* (*plein de signes*) ainsi qu'une erreur avec le figement *près de*. En ce qui concerne la première erreur, il est évident ici que *de signes* est complément de *plein*. Notons qu'ici il n'est pas possible de considérer *plein de* comme un figement total puisqu'il varie en genre (*pleine de*) et en nombre (*pleins de*), au contraire de son homonyme lexicalisé occupant la fonction de déterminant (*plein de frites sont tombées du sachet*). Il est probable que cette prédiction soit influencée par les adjectifs en situation de prédicat dans une structure *xcomp* (→ 4.2.9) ; dans ce cas, les compléments prépositionnels essentiels du prédicat reçoivent cette étiquette.

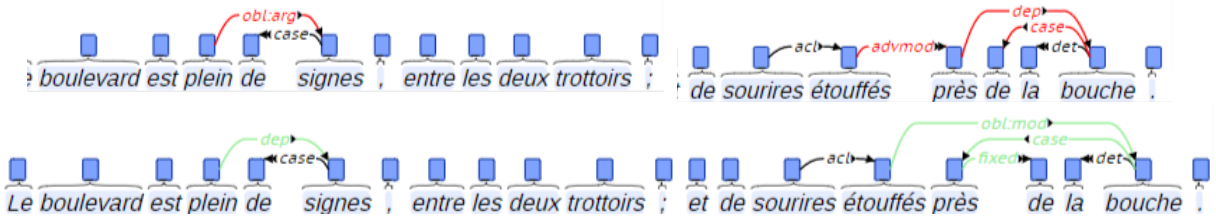


Fig. 50 — Prédiction et correction de PT-4

En ce qui concerne l'erreur de VV-9 (fig. 51), la difficulté qui apparaît est plutôt celle de la sélection du gouverneur d'un complément prépositionnel : rien que dans le syntagme nominal *les élans des rayons dorés sous la paupière […]*, le syntagme prépositionnel *sous la paupière* est susceptible d'occuper la fonction de complément déterminatif (*nmod*) de *élans*, de complément déterminatif de *rayons* (*nmod*) ou de complément prépositionnel de l'adjectif *dorés* (*dep*[52]). C'est par le sens que nous déterminons ici qu'il est plus probable que le complément dépende de *dorés* et qu'il s'agit donc d'une erreur. Il n'est donc pas étonnant de remarquer que les quatre erreurs ambiguës sont de type BOTH et consistent en réalité en une erreur de sélection du gouverneur. Notons que, contrairement à de nombreux cas (→ 4.2.4), le gouverneur

[52] Remarquons qu'UD est ambigu dans le cas des compléments prépositionnels lorsqu'ils dépendent de l'épithète d'un substantif qui pourrait être un adjectif ou un verbe au participe (dans *les élans dorés sous la paupière*, *dorés* pourrait être un adjectif ou le participe passé du verbe *dorer*). En effet, un verbe au participe serait étiqueté par la relation *acl*, indiquant sa nature propositionnelle, et ses compléments prépositionnels recevraient l'étiquette *obl:mod* (« *élans* acl> *dorés* » et « *dorés* obl:mod> *paupière* »). Au contraire, l'adjectif est indiqué *amod* et ses compléments prépositionnels, *dep* (« *élans* amod> *dorés* » et « *dorés* dep> *paupière* »). Lorsque le choix n'est pas évident, nous conservons l'annotation initiale du modèle (*amod* dans le cas de *dorés*) et corrigeons le complément en conséquence. Il s'agit pour nous d'une faiblesse théorique du modèle.

prédit ici est le substantif le plus éloigné alors que le gouverneur correct est le voisin direct du complément.

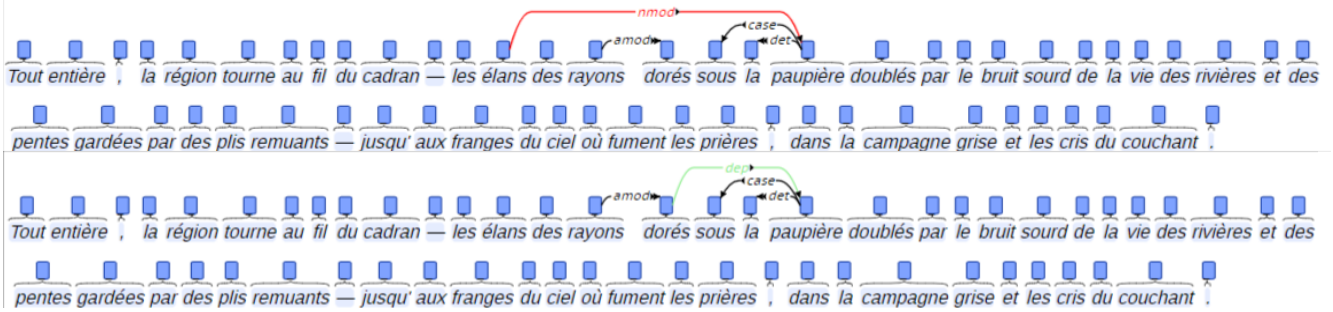


Fig. 51 — Prédiction et correction de VV-9

La même erreur est commise dans PT-4 pour une proposition complément de l'adjectif *impossible* (*impossibles à dire*).

4.2.8 Interrogation : pronom interrogatif et inversion sujet-verbe

Nous regroupons dans cette section deux types d'erreurs puisqu'ils sont coprésents à plusieurs reprises : les erreurs liées au pronom interrogatif[53] *où* (***Où** es-tu ?*) et celles liées à l'inversion de l'ordre du sujet et du verbe (*Que **mange-t-il** ?*), souvent présentes dans les propositions interrogatives directes (Grevisse et Goosse, 2016 : 531). Il n'est pas particulièrement étonnant que nous rencontrions ce type d'erreur en observant le corpus d'entraînement : *où* interrogatif n'est présent que 3 fois dans Sequoia-Train, et nous ne dénombrons que 37 phrases interrogatives directes. Il s'agit donc d'une structure syntaxique que le modèle connaît peu.

Nous avons rencontré ce type d'erreur dans **9 phrases**[54]. Il est à noter que toutes les erreurs de cette section sont incohérentes.

4.2.8.1 *Où* interrogatif

La prédiction de chacune des cinq occurrences dans notre corpus du pronom interrogatif *où*, qu'il soit utilisé seul (G-1, APN-5, EF-2) ou précédé d'une préposition

[53] Contrairement à ce que considèrent Grevisse et Goosse (2018 : 533), classant *où* interrogatif dans la catégorie des adverbes interrogatifs, nous souscrivons plutôt à la conception d'Hadermann (1993 : 39) : « Où, tout comme d'autres mots appelés traditionnellement adverbes, est un pronom ou du moins une proforme. *Où* exprime une indétermination du lieu et a par conséquent un caractère lacunaire qui lui permet de représenter un syntagme (prépositionnel) et d'y référer anaphoriquement, cataphoriquement et même exophoriquement ». Cela est également plus cohérent avec notre méthode d'annotation et celle du corpus d'entraînement, Sequoia : *où* interrogatif est considéré complément oblique essentiel (*obl:arg*), ce qui correspond à un syntagme prépositionnel, ici représenté par le pronom. L'étiquette de complément adverbial (*advmod*) serait plus cohérente en considérant *où* interrogatif comme adverbe.

[54] G-1, APN-5, AA-7, EF-2, NV-10, NV-14, CS-4, VF-8 et RI-4.

(*jusqu'où*, AA-7 ; *d'où*, NV-14), a mené à une erreur DEPREL ; il s'agit donc d'une difficulté que le modèle rencontre, probablement à cause de sa faible fréquence d'apparition dans le corpus d'entraînement.

Ce qui est intéressant dans ce cas, c'est que seules deux étiquettes (erronées) sont utilisées à la place de l'étiquette correcte (*obl:arg*), et leur distribution est systématique : *où* est annoté *dep* lorsque le verbe auquel il se rapporte est à un temps simple et *nsubj* lorsque le verbe est à un temps composé.

Si nous considérons les cas dans lesquels *où* est annoté *dep* (fig. 52), nous constatons en effet que le verbe auquel le pronom interrogatif se rapporte est à un temps simple : *finira* (APN-5) et *poussera* (AA-7), formes à la troisième personne de l'indicatif futur simple de *finir* et *pousser* ainsi que *vient* (NV-14), troisième personne de l'indicatif présent de *venir*.

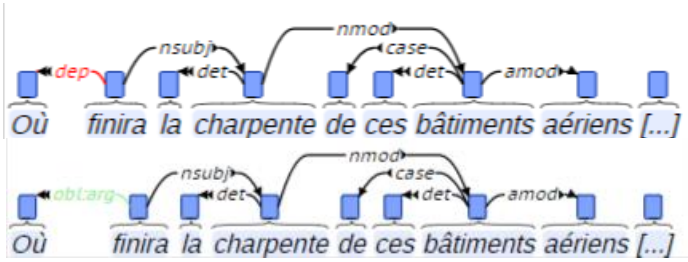


Fig. 52 — Prédiction et correction d'APN-5

Les occurrences étiquetées *nsubj*, apparaissant lorsque les verbes sont à un temps composé, sont plus surprenantes puisque nous obtenons un arbre syntaxique contenant deux tokens *nsubj*, donc sujet nominal, du même verbe (fig. 53).

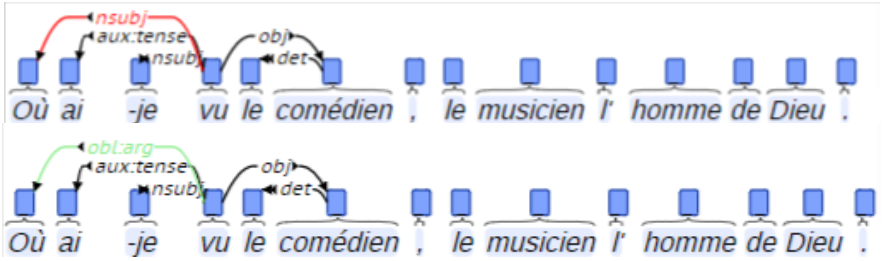


Fig. 53 — Prédiction et correction de G-1

Cela est particulièrement étrange puisque le sujet, malgré l'inversion, est reconnu et il est impossible de réaliser plusieurs fois la fonction sujet en français sans l'usage d'un moyen de coordination, selon le principe d'unicité. Nous pourrions formuler l'hypothèse que le modèle attribue un sujet à chacun des verbes, puisque cette situation n'apparaît que lorsque le temps utilisé est un temps composé. Cependant, le gouverneur d'*où*, du sujet (*-je* dans notre cas) et de l'auxiliaire (*ai*) est bien reconnu dans le

participe passé (*vu*), et cela également dans EF-2 qui présente la même anomalie ; cette hypothèse est donc peu vraisemblable.

Nous ne sommes capables de formuler aucune hypothèse convaincante quant à l'usage de *dep* : dans le corpus d'entraînement, *où* (interrogatif ou relatif) est toujours étiqueté *iobj* ou *obl:arg*. Il est donc improbable que l'étiquette *dep* soit utilisée pour cette unité, si ce n'est en vertu d'une éventuelle récurrence, que nous ne pouvons que supposer puisque nous ne l'avons pas constatée, de tokens *dep* lorsque le sujet ne précède pas le verbe.

4.2.8.2 Inversion sujet-verbe

En ce qui concerne l'inversion sujet-verbe, nous constatons des erreurs dans des constructions dont l'inversion est liée à une interrogation (*me reconnaîtrai-je*, CS-4 ; *que veulent dire cette main…*, RI-4 et *N'y avait-il pas…*, NV-10) ou à une expression, ou du moins une construction exclamative (*Faut-il*, VF-8 ; Grevisse et Goosse, 2016 : 554). Hormis les erreurs concernant RI-4, qui relèvent plutôt de la projectivité (→ 4.2.14.1), il ne s'agit encore une fois que d'erreurs d'étiquette. Aucune régularité n'a pu être remarquée ici.

Les erreurs de CS-4 sont particulièrement incohérentes : le pronom clitique *me* reçoit la fonction de sujet (*nsubj*) alors que, comme le relèvent Grevisse et Goosse (2016 : 919), *me* est la forme des fonctions objet direct, objet indirect et pronom réfléchi ; la forme sujet de la première personne du singulier est nécessairement *je*. Le sujet enclitique, lui, est étiqueté *dep*. Dans ce cas, nous pouvons supposer que l'inversion a joué un rôle déterminant, puisque c'est le pronom qui se trouve à la place traditionnelle du sujet qui reçoit l'étiquette *nsubj*. Le pronom *-je*, qui suit le verbe, ne peut être un des autres actants verbaux puisqu'il n'est pas dans une construction indirecte et recevrait donc une étiquette de dépendance non spécifiée *par défaut*. La raison pour laquelle nous n'obtenons pas de nouveau une duplication de l'étiquette *nsubj* n'est pas claire.

Fig. 54 — Prédiction et correction de CS-4

Dans le cas de *faut-il*, il s'agit d'une confusion entre *expl:subj* et *dep*. La raison pour laquelle, contrairement aux interrogations avec *où*, le sujet enclitique (*-il*) n'est pas compris comme sujet nous échappe, d'autant que la phrase manque alors d'un sujet. Peut-être est-ce dû à la construction impersonnelle, dont *il* n'est que sujet apparent[55] ? L'étiquette concernée est alors *expl:subj*, dont la fréquence est nettement moins élevée dans le corpus d'entraînement. Nous pourrions également considérer que cela est dû à une lexicalisation de la tournure, comme le suggèrent Grevisse et Goosse (2016 : 554) : « On peut aussi considérer comme des locutions servant à introduire une exclamation les formules suivantes [dont *faut-il que* et *faut-il* avec un infinitif], qui ont perdu leur sens originaire ». Dans ce cas, *-il* devrait recevoir l'étiquette *fixed* pour indiquer le figement. Nous n'avons pas choisi cette option en vertu de la tendance d'UD à limiter l'étiquette de figement aux constructions qui ne sont plus analysables, car devenues tout à fait unitaires (→ 4.2.2).

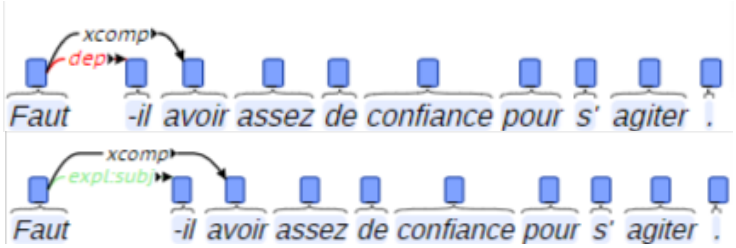


Fig. 55 — Prédiction et correction de VF-8

4.2.9 Homonymie

Comme tout locuteur durant son activité quotidienne, le parser s'est retrouvé face à une grande quantité de formes pouvant représenter différents homonymes (homographes dans notre cas). Lorsque ces homonymes ne sont pas équivalents sur le plan grammatical, c'est-à-dire qu'il existe plusieurs unités partageant un signifiant — sa forme écrite dans notre cas — mais n'appartenant pas à la même catégorie de partie du discours, la désambiguïsation est essentielle pour l'analyse syntaxique. Cette opération peut justement être réalisée grâce à la syntaxe : dans *le rebelle₁ se rebelle₂*, aucune ambiguïté n'est possible ; *rebelle₁*, précédé d'un déterminant, est nécessairement le substantif, et *rebelle₂* est nécessairement une forme du verbe *se rebeller*. Il n'est pourtant pas exclu que le modèle soit, lui, trompé. Au contraire, une phrase comme C-4

[55] Pour plus de précision, nous renvoyons le lecteur à la section consacrée aux verbes impersonnels et présentatifs (→ 4.2.11).

(*Sur la route, mon ombre me suit, oblique, et me dit que je cours trop vite*), le statut de *oblique* n'est pas clair : s'agit-il de l'adjectif *oblique* ou d'une forme du verbe *obliquer* ? Cela a évidemment des conséquences sur l'analyse syntaxique.

Quatre phrases, PA-9 (fig. 56), C-4 (fig. 57), VF-6 (fig. 58) ont retenu notre attention à ce sujet : des erreurs liées à l'homonymie y sont particulièrement visibles.

Le signifiant problématique de PA-9 est *souverain*, qui peut représenter un substantif masculin singulier (*le* ***souverain***$_{subs.}$[56]) ou un adjectif masculin singulier (*un pays* ***souverain***$_{adj.}$). Outre la prédiction du morphologiseur[57], annotant *souverain* comme adjectif, nous pouvons estimer quel homonyme le parser syntaxique a sélectionné par l'étiquette de dépendance *amod*, correspondant à la fonction d'épithète. Or, malgré l'ambiguïté syntaxique possible dans ce cas, il nous semble assez évident, en particulier par l'encadrement entre virgule de *souverain de lui-même*, que *souverain* est un substantif, apposition de *pauvre*. Cela a pour conséquence une erreur DEPREL pour la dépendance entre *souverain* et *pauvre* (*appos* plutôt qu'*amod*), mais aussi une erreur d'étiquette pour les dépendants de *souverain* : lui-même devient alors complément déterminatif (*nmod*). Notons que la dépendance « *souverain* obl:arg> *lui-même* » était erronée dans tous les cas (→ 4.2.7). Il est également intéressant de constater que la virgule a dû jouer un rôle dans la décision malgré l'erreur et n'a donc probablement pas été ignorée : sans elle, *pauvre* est ambigu et l'option *le pauvre*$_{adj.}$ *souverain*$_{sub.}$ *de lui-même* est parfaitement plausible.

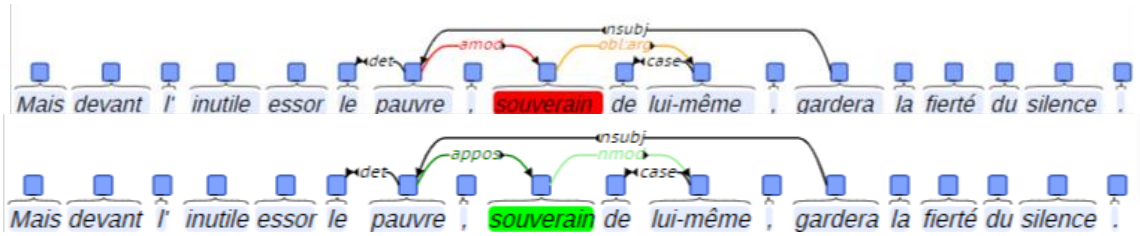


Fig. 56 — Prédiction et correction de PA-9

C-4 (fig. 57) montre une erreur due au signifiant *oblique*, qui peut représenter un adjectif (*ligne* ***oblique***$_{adj.}$), un substantif (*une* ***oblique***$_{subs.}$) ou la forme à la troisième

[56] Pour des raisons de clarté, lorsqu'une ambiguïté est possible en raison d'homonymes, nous indiquons en indice la nature du signe linguistique dont nous parlons.

[57] Nous n'évaluons pas le morphologiseur et nous évitons donc de nous baser sur ses prédictions, mais il peut permettre, de façon ponctuelle, de soutenir des hypothèses. En effet, le morphologiseur est également basé sur les transformers de CamemBERT et la compréhension du contexte est donc très similaire.

personne du présent de l'indicatif du verbe *obliquer* (*il* ***oblique***$_{\textbf{\textit{verbe}}}$). Dans ce cas, la prédiction problématique est « *suit* amod> *oblique* », qui montre que le token a été considéré comme un adjectif puisque le parser lui prédit une étiquette d'épithète (*amod*). Cette prédiction est incohérente (au sens typologique) puisqu'un adjectif épithète ne peut dépendre d'un verbe, bien que la structure aurait pu être ambiguë en considérant *oblique* comme épithète détachée d'*ombre*. Cependant cela aurait rendu la dépendance non projective, ce qui est problématique pour le parser (→ 4.2.14.1). Nous estimons que l'analyse syntaxique la plus cohérente est celle de la coordination, en considérant *oblique* comme verbe : *mon ombre me suit [et] oblique*$_{verbe}$ *[et] me dit que je cours trop vite*. Cette analyse est projective et aisée pour le parser, qui a déjà détecté une coordination à l'aide de la présence de la conjonction de coordination *et*.

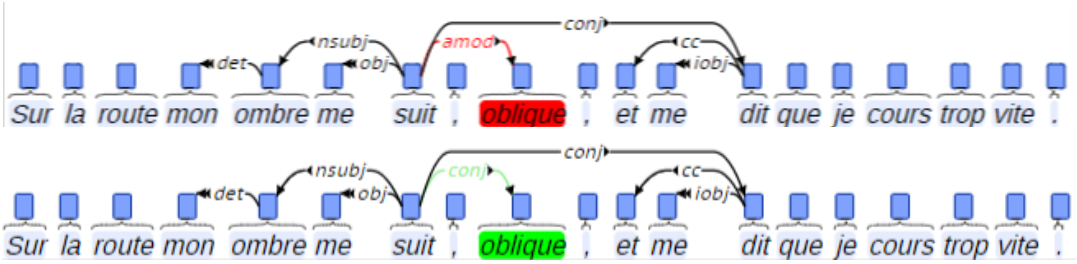


Fig. 57 — Prédiction et correction de C-4

Dans le cas de VF-6 (fig. 58), nous constatons une erreur cohérente due à *sombre*, qui peut être un adjectif (*une pièce* ***sombre***$_{\textbf{\textit{adj.}}}$) ou la troisième personne du présent de l'indicatif de *sombrer* (*il* ***sombre***$_{\textbf{\textit{verbe}}}$). Bien que l'étiquette *advmod* soit compatible autant avec *sombre*$_{adj.}$ que *sombre*$_{verbe}$, le fait que la prédiction n'inclue pas de sujet nous pousse à l'hypothèse de l'adjectif, ce qui est renforcé par la prédiction du morphologiseur (*ADJ*). Pour l'analyse syntaxique de la phrase, l'hypothèse verbale (*sombre*$_{verbe)}$ est plus convaincante à nos yeux : la conjonction de coordination *et* coordonne alors deux syntagmes verbaux et il n'est pas nécessaire de supposer l'élision d'un verbe et de son sujet pour la première partie de la phrase (*tout sombre*), qui n'est pas autonome. *Tout*, qui est un signifiant hautement homonymique lui aussi, est alors un pronom indéfini et a pour fonction d'être le sujet de *sombre*$_{verbe}$.

Fig. 58 — Prédiction et correction de VF-6

L'homonymie est donc bien sûr problématique pour le parser, comme elle l'est parfois dans la communication : la résolution de l'ambiguïté passe souvent par la prise en compte attentive du contexte, autant syntaxique que sémantique, ce qui est plus difficile techniquement. Dans certains cas, comme *oblique* de C-4, il est même impossible de trancher avec certitude et le jeu stylistique de la littérature, en particulier la poésie, est susceptible de faire apparaître des structures syntaxiques volontairement ambiguës. L'ambiguïté, qu'elle soit purement syntaxique ou liée à l'homonymie restera donc toujours un obstacle au parsing syntaxique, et à l'annotation en général.

4.2.10 Fonction *xcomp*

La fonction *xcomp*, reprise par UD, est une fonction courante empruntée à la *Lexical Functional Grammar* :

> The COMP [*ccomp* pour UD], XCOMP, and XADJ [également *xcomp* pour UD] functions are clausal functions, differing in whether or not they contain an overt SUBJ noun phrase internal to their phrase. The COMP function is a closed function containing an internal SUBJ phrase. The XCOMP and XADJ functions are open functions that do not contain an internal subject phrase; their SUBJ must be specified externally to their phrase. (Dalrymple, 2001 : 24)

Il s'agit donc de compléments prédicatifs (généralement des propositions à un mode nécessairement non fini [*Pierre veut* ***manger une pomme***], mais également des adjectifs [*Je le considère* ***honnête***]) dont la fonction sujet n'est réalisée qu'à l'extérieur de leur proposition et qui transite par l'intermédiaire du verbe ou d'un semi-auxiliaire (Grevisse et Goose, 2016 : 1137). En effet, dans *Pierre veut manger une pomme*, le sujet de *manger* est *Pierre*, qui est réalisé dans la proposition *Pierre veut*, c'est-à-dire à l'extérieur de la proposition *manger une pomme*. Au contraire, évidemment, les compléments internes restent des dépendants du verbe de la proposition *xcomp* (*une pomme* est bien le complément d'objet direct de *manger*). La structure en dépendance est donc scindée puisque les actants ne dépendent plus tous du même mot : dans le cas d'une construction constituée d'un semi-auxiliaire suivi d'un infinitif, le sujet dépend du semi-auxiliaire et les compléments d'objet de l'infinitif (fig. 59).

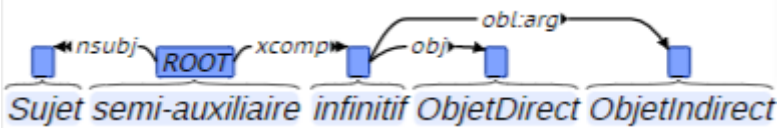


Fig. 59 – Annotation de la fonction *xcomp* dans UD

Dans ce cadre, nous constatons **4 erreurs** dans notre corpus, bien que la plupart des fonctions *xcomp* soient correctement prédites. Deux de ces erreurs, comprises dans EF-5 et TPB-7 nous semblent relever plutôt d'un problème de projectivité apparaissant dans le cas de fonctions *xcomp* à l'intérieur d'une proposition relative dont l'antécédent est complément d'objet direct et ont été traitées séparément (→ 4.2.14.1). En revanche, VF-8 (fig. 60) et NV-1 (fig. 61) font apparaître des erreurs HEAD pertinentes dans cette section.

Pour VF-8, il s'agit d'une proposition adverbiale non essentielle (*obl:mod*), relevant de la phrase principale, qui est considérée comme dépendante du verbe de la fonction *xcomp*. La structure syntaxique est pour autant compliquée : il s'agit d'une fonction *xcomp* dépendant d'un verbe impersonnel, qui n'a qu'un sujet apparent, et c'est donc la proposition *xcomp* elle-même qui est sujet réel de la phrase.

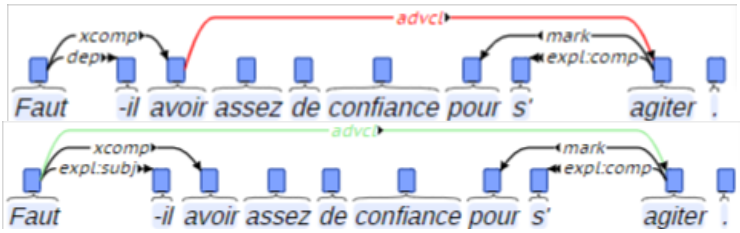


Fig. 60 — Prédiction et correction de VF-8

L'erreur de NV-1 est plutôt inattendue, puisque l'objet (ici une proposition complétive, *ccomp*), dépend du semi-auxiliaire plutôt que du verbe de la fonction *xcomp* alors que les actants, à l'exception du sujet, doivent dépendre de ce verbe. Cependant, l'insertion de *en comptant bien*, complément adverbial de la complétive, entre la conjonction de subordination et le noyau de la proposition peut avoir constitué une difficulté. Dans la prédiction, *qu'* est ainsi considéré comme la conjonction de subordination du complément adverbial, et elle est donc liée au verbe disponible le plus proche (le gérondif *en comptant*). Notons qu'au contraire le gouverneur de la complétive est prédit au verbe le plus éloigné, ce qui montre que le modèle ne minimise pas systématiquement la longueur des dépendances.

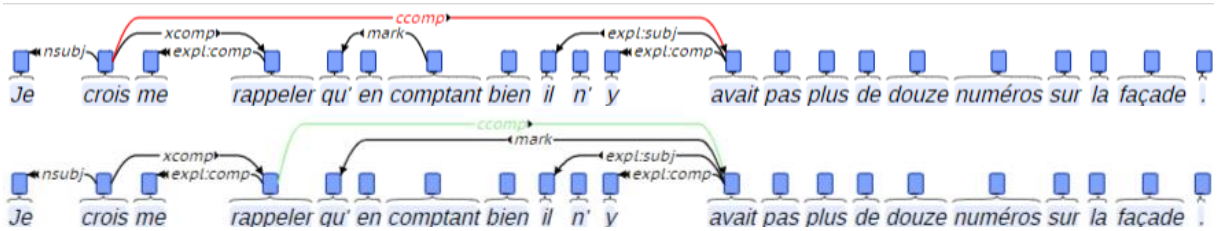


Fig. 61 — Prédiction et correction de NV-1

4.2.11 Propositions à sujet explétif : verbes impersonnels et présentatifs

Nous rassemblons ici toutes les erreurs liées à des constructions dont le sujet est explétif (étiqueté *expl:subj*), c'est-à-dire les propositions à verbe impersonnel ou dont l'introducteur a une fonction de présentatif (Grevisse et Goosse, 2016 : 1529). Ces deux cas sont en effet très proches dans UD et nous les considérons conjointement.

En grammaire traditionnelle, les verbes impersonnels sont caractérisés par un usage à la troisième personne (Grevisse et Goosse, 2016 : 1097) et la présence d'un sujet dit *grammatical* (ou *apparent*), généralement *il* ou *ça*, ainsi qu'occasionnellement un sujet *logique* (ou *réel*) (Grevisse et Goosse, 2016 : 270), construit comme un complément. Pour des raisons d'uniformité, le sujet réel est analysé comme le complément d'objet direct dans le formalisme UD (c'est-à-dire que la tête du complément dépend du verbe selon l'étiquette *obj*). Le sujet apparent est, lui, remarqué : on considère qu'il est explétif et il reçoit donc l'étiquette de sujet explétif *expl:subj* (fig. 62).

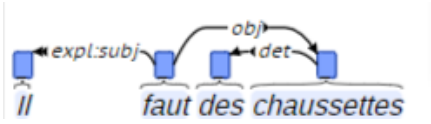


Fig. 62 — Annotation du syntagme d'un verbe impersonnel

Le présentatif *il y a* a un fonctionnement tout à fait semblable aux verbes impersonnels et est traité similairement : par exemple, dans SA-5[58], le sujet (*il*) est considéré comme explétif et le complément (*des gens venus de partout et qui parlent*), que nous appelons ici *complément du présentatif*, reçoit un traitement similaire au complément d'objet direct (fig. 63).

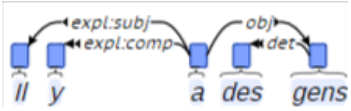


Fig. 63 — Arbre syntaxique d'*il y a des gens*

Le présentatif *c'est*, ainsi que le tour impersonnel *il est question de* sont cependant analysés comme des constructions avec copule (→ 4.2.12.1), ce avec quoi nous entrons en désaccord au motif que, dans *il est question*, *il* n'a plus aucune valeur sémantique et ne peut constituer un sujet auquel se rapporte un attribut, et *question* ne semble

[58] **Il y a** des gens venus de partout et qui parlent — les têtes ramenant l'esprit qui se souvient — et le ciel, qui descend plus lourd sur l'arbre qui se dresse, ouvre une porte basse par où tombe le soir.

aucunement attribut. Dans le cas de *c'est*, la situation est plus délicate puisque la version lexicalisée comme présentatif *c'est*, dans laquelle *c'* est un sujet explétif (voir AA-6[59]), mériterait d'être analysée comme *il y a*. Au contraire, lorsque *c'* reste un pronom anaphorique (voir AP-2[60] ; pour nous, la forme au singulier ne vient pas d'un figement, mais de l'accord sémantique avec le complément du présentatif), il doit être considéré sujet nominal (*nsubj*) qualifié par un attribut du sujet à l'aide d'une copule. Cependant, la distinction est assez floue et dépend de l'interprétation de l'annotateur, ce qui constitue un problème. Nous pouvons donc comprendre que Sequoia considère systématiquement le pronom *c'* comme *nsubj*, à l'exception des tours emphatiques de type *c'est… qui*. Nous ne pouvons donc raisonnablement pas considérer ceci comme des erreurs du parser.

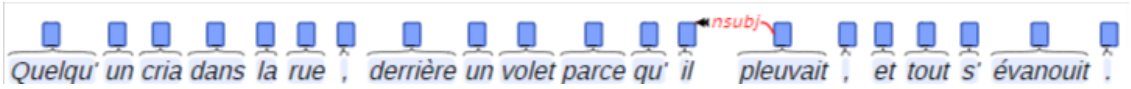


Fig. 64 — Prédiction de G-6

La plupart des 28 étiquettes *expl:subj*, sujets des verbes impersonnels, sont prédites correctement. Nous n'avons constaté des erreurs que dans **4 phrases** : C-9 (*il est question*), G-6 (*il pleuvait*, fig. 64), VF-8 (*il faut*) et NV-10 (*il y a*). Dans les deux premières, l'étiquette *nsubj* est substituée à *expl:subj* ; dans les deux secondes, il s'agit de l'étiquette *dep*, ce dont nous parlons dans la section des erreurs d'inversion (→ 4.2.1). Notons qu'il s'agit toujours de faux négatifs.

4.2.12 Erreurs liées à des structures conventionnelles d'UD

Les erreurs de cette section sont, pour nous, liées à des particularités des règles d'annotation UD. Il s'agit d'erreurs liées à la construction à verbe copule (→ 4.2.12.1), aux verbes factitifs (→ 4.2.12.2), à la comparaison relative (→ 4.2.12.2) et aux phrases clivées (→ 4.2.12.4).

4.2.12.1 Construction à verbe copule

Comme la plupart des linguistes, les règles d'annotation UD différencient l'attribut du sujet d'un actant comme le complément d'objet direct, et effectue une distinction entre les constructions dans lesquelles le verbe est la tête du prédicat (« verbes

[59] **C'est** la voix de la foule obscure qui murmure ou le bruit des pas qui battent le chemin.
[60] Ces lignes qui filent dans le creux sans fin et se rejoignent **c'est** la route de l'imagination sans surprise.

d'action[61] » ; *Pierre **mange une pomme***) et celles dans lesquelles le verbe fait partie du prédicat (copules[62] ; Pierre **est un humain**) (Grevisse et Goosse, 2016 : 1067). Cela est intégré dans l'arbre syntaxique par le choix de la tête du prédicat dans le cas des attributs du sujet : dans les constructions à copule, c'est l'attribut du sujet qui est la tête de la proposition et la copule dépend de lui (fig. 65). De même, tout complément de phrase, dépendant de la tête de la proposition, devient alors dépendant de l'attribut plutôt que du verbe. Nous comprenons donc qu'il est aisé, dans cette structure renversée, que des erreurs apparaissent lorsque la copule n'est pas considérée.

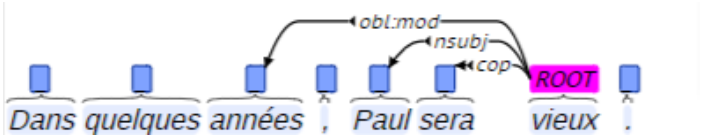


Fig. 65 – Construction à verbe copule selon UD

20 erreurs, réparties en 9 phrases[63] sont à nos yeux liées aux constructions à copule. Les erreurs sont dues à trois facteurs dans notre cas : la présence d'une copule moins commune qu'*être* (*rester* et *avoir l'air* [Grevisse et Goosse, 2016 : 288]), une expression semi-figée (*être en train de*) ou l'ellipse de la copule.

a) Copule moins fréquente

Dans cet exemple (fig. 66), *restait* est analysé comme un verbe d'action ; *indécise* est donc prédit incorrectement. Cette simple inversion conduit directement à trois erreurs BOTH.

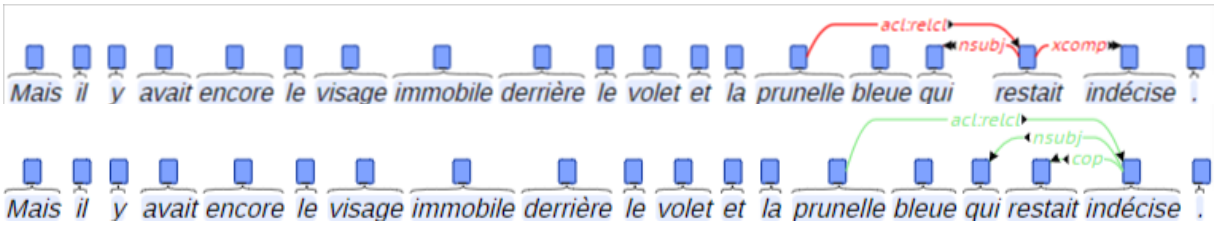


Fig. 66 — Prédiction et correction de NV-6

[61] Cette terminologie, rencontrée notamment chez Tesnière (1959), est assez pauvre, mais pratique dans cette section pour marquer l'opposition entre les structures à attribut du sujet et les autres types de prédicats. Cela rassemble notamment les verbes transitifs, les verbes intransitifs et les verbes réfléchis.

[62] Nous nous rangeons ici à l'opinion de Grevisse et Goosse (2016 : 285), qui désignent par *copule* tous les verbes unissant le sujet et l'attribut du sujet, bien que ceux-ci notent que certains linguistes préfèrent utiliser le terme *verbe attributif* pour tout verbe autre qu'*être* et occupant cette fonction.

[63] C-1, C-2, C-5, C-8, AR-1, TE-1, LPN-3, NV-6 et AP-8.

b) Expression semi-figée

La figure 67 présente le cas d'expression semi-figée que nous rencontrons. Puisqu'UD décompose analytiquement toutes les expressions qui ne sont pas complètement lexicalisées, *être en train de* est considéré comme une construction avec une copule. Le fait qu'elle ne soit pas reconnue comme telle entraîne encore une fois des erreurs en série.

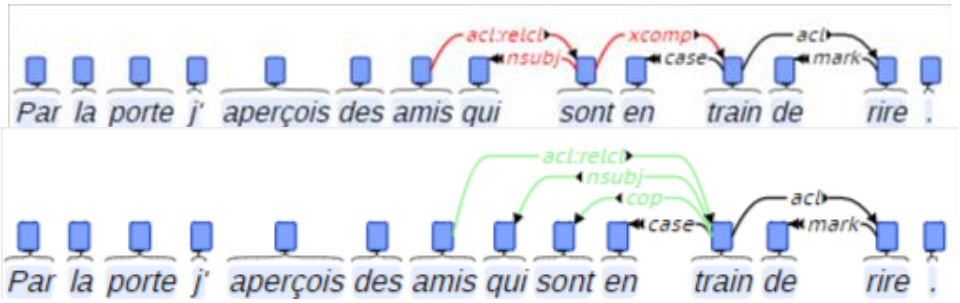


Fig. 67 — Prédiction et correction de C-8

c) Cas d'ellipse

Les cas d'ellipse de copule sont traités dans la section consacrée aux ellipses (→ 4.2.3).

Le fait qu'UD déplace la tête du prédicat du verbe à l'attribut lorsqu'il s'agit d'une construction attributive ne semble pas être particulièrement problématique tant que la copule est reconnue : nous avons de nombreux cas dans le corpus dans lesquels la prédiction est parfaite. Cependant, lorsque des verbes copules plus rares sont utilisés ou lorsque la copule est effacée, les erreurs s'enchaînent : à la place d'avoir une simple erreur d'étiquette de la dépendance du verbe, les prédictions du gouverneur de tous les actants et circonstants sont erronées. Un paradigme d'annotation dans lequel la structure du prédicat est plus stable diminuerait donc le nombre d'erreurs de ce type.

4.2.12.2 Verbes factitifs

Dans le corpus prédit, nous obtenons 4 erreurs réparties en 2 phrases (FP-1 et NV-8) liées à une structure dite *causative* dans UD, c'est-à-dire en présence de verbes factitifs, en particulier le semi-auxiliaire *faire*.

> Un verbe factitif (on dit aussi *causatif*) est un verbe dont le sujet fait faire l'action exprimée par le verbe. […] Le verbe faire suivi d'un infinitif est un semi-auxiliaire transformant n'importe quel verbe en verbe factitif, même *faire* lui-même. (Grevisse et Goosse, 2016 : 1076)

Dans ce cas, il est conventionnel dans l'environnement UD de considérer *faire* comme un auxiliaire classique, dépendant de l'infinitif qui le suit selon l'étiquette *aux:caus*, tout en sélectionnant l'infinitif comme tête verbale. La fig. 68 illustre cette construction et les erreurs rencontrées.

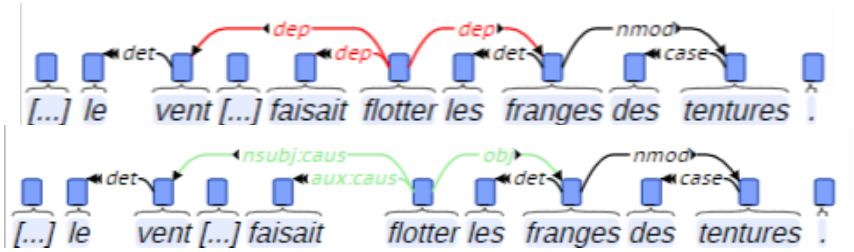


Fig. 68 — Prédiction et correction de NV-8

Il est surprenant dans NV-8 que la structure soit reconnue, plaçant effectivement *flotter* comme tête de la proposition, sans que les étiquettes correctes soient prédites. En effet, les structures à semi-auxiliaire classiques (*xcomp*, 4.2.9) considèrent que le semi-auxiliaire est gouverneur du sujet et que l'infinitif dépend du semi-auxiliaire selon l'étiquette *xcomp* ; nous aurions pu nous attendre à ceci puisque nous rencontrons encore une fois l'enchaînement d'un sujet, d'un verbe à un mode fini et d'un verbe à l'infinitif. La structure causative est donc reconnue par le parser, mais il n'a néanmoins pas prédit les étiquettes correctes (*nsubj:caus* et *aux:caus*, spécifiques aux constructions avec *faire*, et *obj*), qui sont alors sous-spécifiées. Puisque *nsubj:caus* et *aux:caus* sont des étiquettes rares, nous pouvons lier la difficulté au manque d'entraînement. Le fait que l'étiquette *obj*, courante, ne soit pas prédite alors qu'il est très régulier — c'est-à-dire dans une majorité de fonctions *xcomp* — qu'un infinitif soit gouverneur d'un token *obj* est plus étonnant.

Notons que l'étiquette *aux:caus* n'est rencontrée que 18 fois dans le corpus d'entraînement (Sequoia-Train), et que les deux structures rencontrées dans notre corpus sont prédites incorrectement du point de vue de l'étiquette.

4.2.12.3 Comparaison relative

La comparaison (ou comparatif [Grevisse et Goosse, 2016 : 1325-1326]) est un point de discussion dans l'environnement UD et fait l'objet d'un groupe de travail[64]. Il

[64] https://universaldependencies.org/workgroups/comparatives.html.

n’est donc pas étonnant que des erreurs apparaissent autour de cette structure très particulière.

Fig. 69 — Comparaison de l’attribut (gauche) et du prédicat (droite)

Prenons l’exemple simple *David est plus rapide que Charles*. Il s’agit ici d’un comparatif de supériorité de construction analytique (Grevisse et Goosse, 2016 : 1326) exprimé par l’intermédiaire de l’adverbe de degré *plus*. La phrase compare ainsi un comparé qui est l’entité déterminée, *David* avec un standard de comparaison permettant de donner un point de comparaison, *Charles*. Nous nommons *paramètre de comparaison* l’adjectif ou adverbe autour duquel la comparaison est formée ; il s’agit de *rapide* dans notre cas. Puisque *rapide* qualifie *David*, il s’agit de son dépendant. De même, puisque l’adverbe de degré agit sur le paramètre de comparaison, il s’agit du dépendant de ce dernier. Enfin, le standard de comparaison (dans notre cas, *que Charles*) est inclus ordinairement dans une proposition conjonctive corrélative souvent elliptique par suppression des éléments déjà exprimés (Grevisse et Goosse, 2016 : 1329), qui modifie elle aussi le paramètre de comparaison et est donc également son dépendant. Nous obtenons alors une structure du type de la figure 69, centrée autour du paramètre de comparaison. Puisque toutes ces fonctions sont adverbiales, nous leur assignons des étiquettes appropriées.

Lorsqu’il ne s’agit pas d’une construction à copule, il est également possible que le comparé soit un prédicat (*David grimpe plus vite que Charles* ; fig. 69). Le standard de comparaison est encore une fois en réalité une proposition conjonctive corrélative dans laquelle le verbe est omis afin de ne pas le répéter (*David grimpe plus vite que Charles [grimpe]*).

Cette structure complexe entraîne au moins **4 erreurs** dans 2 phrases. Nous pourrions ajouter PA-1, contenant également un comparatif, mais les erreurs concernées sont indissociables des difficultés de choix de la racine, de projectivité et de déplacement.

(PA-1) Tel que tu es je t’aime tout à fait, disait-il.

Nous pouvons constater dans NV-2 (fig. 70) que la structure comparative n'est pas du tout reconnue adéquatement. Tout d'abord, le centre de la structure est ici l'adverbe de degré, ce qui cause une erreur pour l'annotation de *à droite*, considéré comme complément prépositionnel d'un adverbe. Cette confusion peut être due à la forme prépositionnelle du paramètre de comparaison (*à droite*) à cause de sa nature de substantif, ce qui expliquerait l'étiquette *dep* (→ 4.2.7). La proposition conjonctive corrélative semble également considérée complément de *plus*. Enfin, le comparé n'est pas correct : dans ce cas, il ne s'agit pas d'une comparaison modifiant un prédicat (*revois le deux* dans son ensemble), mais bien le substantif *deux*¸ selon un fonctionnement de comparaison à copule.

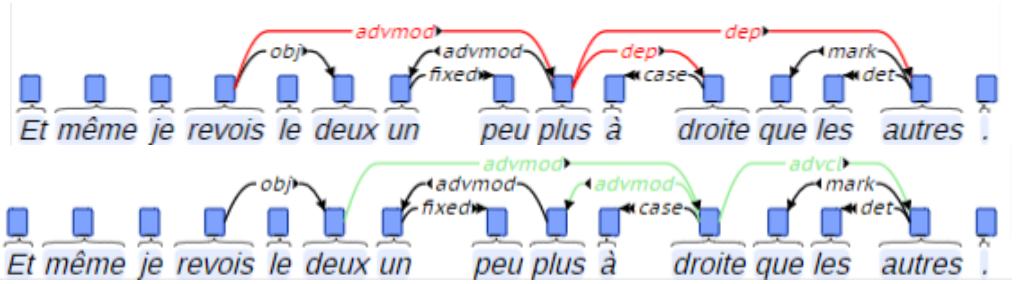


Fig. 70 — Prédiction et correction de NV-2

Dans HE-3 (fig. 71), les annotations du comparé, du paramètre de comparaison et de l'adverbe d'intensité sont correctes. La proposition conjonctive corrélative est cependant considérée encore une fois comme complément prépositionnel d'un adverbe, ce qui est insuffisant dans le cas du comparatif.

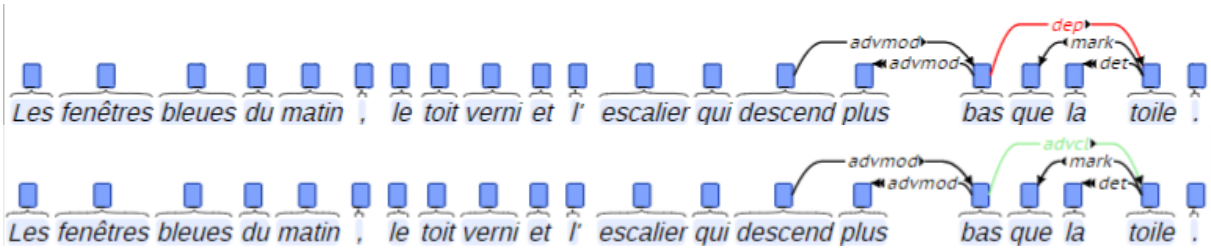


Fig. 71 — Prédiction et correction de HE-3

Dans ce cas, la prédiction erronée est influencée par des paramètres linguistiques, mais aussi par des conventions UD. En effet, nous avons pu remarquer que le comparatif se construit selon une structure linguistique dense centrée autour du paramètre de comparaison et il est important que chacun des éléments constitutifs soit reconnu pour que l'annotation soit correcte. De plus, la typologie des comparatifs, *a fortiori* en incluant les superlatifs, est particulièrement diverse et il est donc

extrêmement difficile d'en dégager un squelette prototypique. Afin de garantir l'annotation cohérente à travers les différents corpus, l'environnement UD a donc dû définir une série de standards plutôt conventionnels pour le comparatif s'opposant par moment avec les règles classiques.

4.2.12.4 Phrases clivées

Trois phrases de notre corpus, C-5, AP-10 et AP-11, contiennent un énoncé clivé (Grevisse et Goosse, 2016 : 627) permettant la mise en relief du sujet (de type *c'est Pierre qui part*) ; aucune des structures n'a été prédite correctement par le parser. À l'exception de C-5, dont les erreurs sont fortement influencées par la difficulté posée par *avoir l'air* (→ 4.2.12.1), il ne s'agit que d'erreurs DEPREL. Cela n'est pas étonnant : du point de vue syntaxique, la seconde partie des énoncés clivés est très similaire aux propositions relatives par la présence du pronom relatif. Le parser, habitué à établir des liens entre un verbe précédé d'un pronom relatif et un substantif ou pronom servant d'antécédent au pronom, prédit alors correctement le gouverneur syntaxique. Cependant, puisque, dans Sequoia, la relation *advcl:cleft* est utilisée exclusivement pour les énoncés clivés, celle-ci est assez peu fréquente et donc difficile à intégrer pour le modèle. Il est donc attendu que le modèle ne prédise pas l'étiquette correcte, mais plutôt une étiquette habituelle dans ces cas : la relation de proposition relative *acl:relcl*. C'est la prédiction du modèle pour AP-10 (fig. 72).

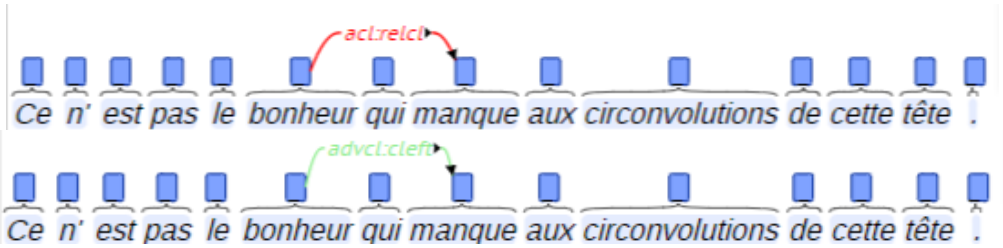


Fig. 72 — Prédiction et correction de AP-10

Il est plus étonnant que la prédiction d'AP-11 (fig. 73) fasse apparaître l'étiquette de relation sous-spécifiée *dep*, alors que la phrase répond à la précédente dans une structure syntaxique exactement similaire (mis à part le fait que les deux syntagmes nominaux *les idées lugubres* et *le nouveau prophète* sont augmentés par l'adjonction d'une épithète). Nous ne pouvons formuler aucune hypothèse sur cette particularité.

Fig. 73 — Prédiction et correction de AP-11

Pour l'annotation syntaxique automatique, la difficulté de ce type de construction est due à sa ressemblance, en contexte réduit, avec la proposition relative ainsi qu'à la faible fréquence de la relation *advcl:cleft* dans le corpus d'entraînement (12 occurrences). Il est tentant de critiquer UD pour la création d'une étiquette utile uniquement à ce type d'énoncé minoritaire fonctionnant syntaxiquement comme un présentatif dont le complément est modifié par une proposition relative, d'autant que l'étiquette *advcl:cleft* n'est présente que dans les corpus du français et du naija[65]. Nous déplorons néanmoins, en tout cas en nous focalisant sur le français, que la relation choisie soit *advcl:cleft* : pour nous, le complément contenant le pronom relatif (par exemple *qui échappe aux idées lugubres de ce nouveau prophète* dans AP-11) n'est pas une proposition adverbiale (*advcl*), mais plutôt une proposition ayant la fonction de complément d'un nom (*acl*), dans notre cas *grandeur*, comme l'est la proposition relative. Cette dernière est étiquetée avec le sous-type *acl:relcl* ; il est alors pertinent, pour nous, que la relation de la phrase clivée utilise un nouveau sous-type de la relation *acl* : *acl:cleft*.

4.2.13 Structures uniques et erreurs exceptionnelles

Cette section comprend trois catégories d'erreurs qui sont soit exceptionnelles pour le parser et la seule hypothèse explicative plausible est le hasard, soit liées à une structure n'apparaissant qu'une fois dans notre corpus et de laquelle, tel un hapax, nous ne pouvons donc déduire aucun comportement particulier. Les erreurs de ponctuation (→ 4.2.13.1) et de dépendance d'auxiliaire (→ 4.2.13.2) entrent dans la première catégorie. Les erreurs de discours rapporté (→ 4.2.13.3) et d'interjection (→ 4.2.13.4) sont uniques, car ces éléments n'apparaissent qu'une fois dans notre corpus ; il est cependant probable qu'en présence d'une plus grande quantité de ceux-ci, les erreurs eussent été plus nombreuses.

[65] https://universaldependencies.org/fr/dep/advcl-cleft.html.

4.2.13.1 Analyse de la ponctuation

Bien que ce type d'erreur soit extrêmement rare, nous avons jugé bon de leur donner une place à part dans notre exposé, sans tenter de les intégrer à une section plus générale. En effet, la ponctuation elle-même a une place particulière dans la grammaire de dépendance : une marque de ponctuation (dans notre cas un token représentant une marque de ponctuation, ci-après token-ponctuation, par opposition à token-mot) ne peut avoir que *punct* comme étiquette et est nécessairement une feuille (c'est-à-dire un élément terminal, n'ayant pas de descendant) de l'arbre syntaxique. Les différentes méthodes d'annotation en disposent librement : certaines, comme la nôtre, l'ignorent, d'autres, comme UD[66], établissent une série de principes soumis à exception dès lors que cela peut créer une dépendance non projective (→ 4.2.14.1). Ces erreurs permettent d'illustrer le problème que nous pourrions appeler « de cécité » des réseaux de neurones : pour le modèle, les données en entrée et les prédictions n'ont aucun sens et ne correspondent qu'à des nombres ; le modèle est donc incapable de toute réflexivité ou prise de distance vis-à-vis des prédictions. Le modèle est en quelque sorte aveugle. Peu importe la qualité du modèle, il sera toujours possible de voir apparaître ce type d'erreur qui semble absurde pour l'œil humain.

Fig. 74 — Prédiction de VF-5

Nous pouvions imaginer a priori que les caractéristiques de la ponctuation ont été tout à fait intégrées par le modèle et qu'aucune erreur ne serait commise, ni avec l'étiquette *punct*, ni dans la prédiction des étiquettes des tokens-ponctuation, mais la vérification empirique nous a donné tort : les phrases VF-5 (fig. 74), RI-1 (fig. 75) en contiennent chacune une et TPB-15 (fig. 76) en contient même deux.

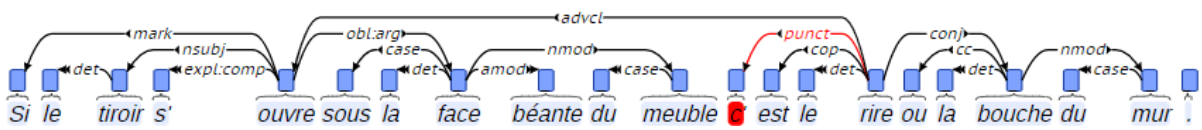


Fig. 75 — Prédiction de RI-1

[66] Rappelons que nous suivons les règles d'annotation UD, mais que nous avons décidé d'ignorer la ponctuation (→ 2.3.2.2).

Dans les deux cas précédents, l'erreur est du même type : il s'agit de tokens-mots, l'adverbe *partout* et le pronom *c'*, qui sont étiquetés *punct*, comme seuls le sont les tokens-ponctuation. Rien ne nous permet de cerner un peu mieux ces erreurs qui semblent être apparues aléatoirement puisque les tokens-ponctuation des deux phrases sont annotés correctement et cette erreur ne peut donc être une conséquence d'une erreur précédente. Nous pouvons simplement relever qu'il ne s'agit dans ce cas que d'erreurs DEPREL.

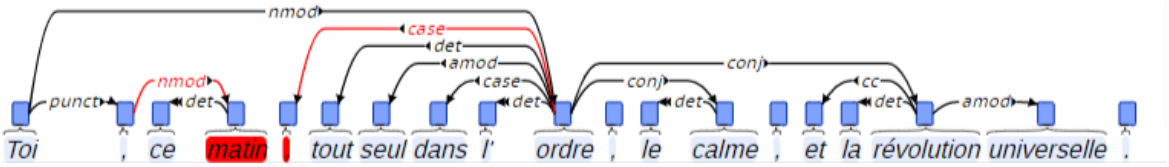


Fig. 76 — Prédiction de TPB-15

TPB-5 montre des erreurs d'un type différent : la première virgule n'est pas une feuille, mais reçoit un dépendant (« , nmod> *matin* ») et la seconde reçoit *case* comme étiquette. Nous avons donc ici un token-mot dépendant d'un token-ponctuation et un token-ponctuation dont l'étiquette n'est adaptée qu'à des tokens-mot. Au vu de l'étrangeté de ces erreurs qui, de plus, se produisent à proximité l'une de l'autre (les erreurs sont liées aux tokens 4 et 5), nous pouvons supposer qu'elles sont liées. Ces erreurs permettent d'illustrer le fait que le comportement d'un modèle statistique ne peut en pratique pas être cerné, car chaque prédiction est le résultat d'une grande quantité de facteurs qu'un opérateur humain ne pourra jamais comprendre entièrement.

4.2.13.2 Dépendance d'auxiliaire de temps composé

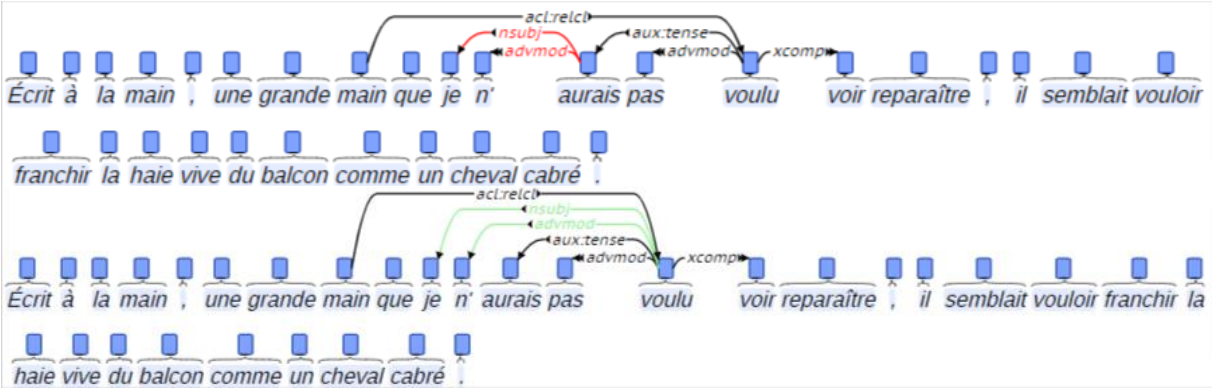


Fig. 77 — Prédiction et correction de NV-3

La phrase NV-3 (fig. 77) présente **une erreur** peu commune : malgré la prédiction correcte de *aurais* comme auxiliaire de temps, le sujet et la particule de négation[67] en dépendent alors que l'auxiliaire de temps est normalement toujours une feuille de l'arbre syntaxique, dépendant du verbe au participe passé. Puisque ce phénomène n'apparaît qu'une fois, malgré de nombreuses prédictions correctes dans d'autres phrases (20 occurrences), nous ne pouvons proposer aucune hypothèse d'explication.

4.2.13.3 Discours rapporté

PA-1 (fig. 78) constitue **un cas unique** dans notre corpus de discours direct (Grevisse et Goosse, 2016 : 567). Selon UD, le discours rapporté avec le verbe *dire* s'analyse grâce à *ccomp*[68] (proposition complément d'objet direct ; *je crois **que tu as raison***) puisque le discours s'identifie comme actant de ce verbe. Cependant, les consignes se contredisent en partie : dans la description de l'étiquette *parataxis*[69], un exemple indique que *John said*, dans *The guy, John said, left early in the morning*, doit être analysé comme parataxe. Bien que, dans le second cas, le verbe de discours soit en incise[70], il nous semble qu'il s'agit là d'une incohérence théorique dommageable pour la prédiction du discours rapporté par un parser.

Dans notre cas, il semblerait que cette confusion ne soit pas la cause de l'erreur : le verbe de discours est prédit comme second membre d'une coordination avec *es*. La difficulté réside plutôt dans le choix de la racine (→ 4.2.2), en bleu dans la figure. Il est cependant très courant, dans le cas du discours rapporté, que le verbe de discours soit placé en fin de phrase, ce qui rend malaisée l'analyse du discours comme proposition complément de ce verbe.

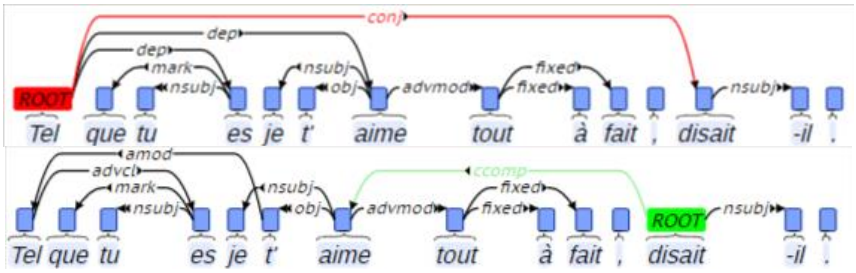


Fig. 78 — Prédiction et correction de PA-1

[67] Il est à noter que la prédiction du gouverneur de *n'* est erronée, mais pas celle de *pas*.
[68] https://universaldependencies.org/u/dep/ccomp.html#reported-speech
[69] https://universaldependencies.org/u/dep/parataxis.html
[70] « When a speech verb interrupts reported speech content, the interruption is treated as a parenthetical parataxis », https://universaldependencies.org/u/dep/parataxis.html#reported-speech

4.2.13.4 Interjection

L'interjection exclamative *ô* (« onomatopée » selon Grevisse et Goosse [2016 : 230]) apparaît dans TPB-5 (fig. 79) et fait l'objet d'une erreur. Celle-ci s'explique probablement par le fait que la forme est extrêmement peu fréquente en dehors de la poésie lyrique et que l'étiquette correcte, *discourse* est rare elle aussi : elle n'est jamais prédite, apparaît 1 fois dans la correction et seulement 3 fois dans Sequoia-Train, accompagnant les interjections *hélas* et *eh*. Un parser entraîné sur base d'un corpus oral serait probablement beaucoup plus performant dans ces cas en raison de la fréquence élevée d'apparition d'interjections dans la conversation.

Fig. 79 — Prédiction et correction de TPB-5

4.2.14 Erreurs liées à des paramètres techniques du parser

Parmi les erreurs, deux types semblent plutôt dépendre de caractéristiques techniques ou de configurations du parser : il s'agit des erreurs de projectivité (→ 4.2.14.1) et de segmentation interne de la phrase (→ 4.2.14.2).

4.2.14.1 Projectivité de l'arbre syntaxique

La projectivité est, comme l'explique Guy Perrier (2021 : 46), une propriété d'un arbre syntaxique qui implique qu'aucune des dépendances syntaxiques ne se croise lorsqu'il est représenté de façon linéaire, c'est-à-dire lorsque ses nœuds sont placés, dans l'ordre de la phrase, sur la même ligne. La fig. 80 donne un exemple des quatre configurations possibles de croisement des dépendances.

La proportion de ces arbres non projectifs dans les corpus dépend de trois paramètres :

- La langue : Gómez-Rodríguez *et al.* (2018 : 2664) expliquent que, dans les corpus annotés selon le format UD, la proportion grimpe jusqu'à 63 % pour le grec ancien, avec une moyenne de 12 % pour l'ensemble des langues.
- Le format d'annotation : selon Perrier (2021 : 43), « le format d'annotation syntaxique joue un rôle important dans l'existence ou non de dépendances non projectives » ; il ajoute que le format SUD (Gerdes *et al.*, 2018) est beaucoup plus propice à l'apparition de la non-projectivité que le format UD.

- Le genre du texte : en effet, il est courant de voir apparaître des dépendances non projectives lors de transformations syntaxiques comme le déplacement d'éléments (par exemple l'apposition ou l'épithète détachées) ou l'insertion au sein d'un syntagme d'un segment dépendant d'un autre gouverneur que la tête de syntagme, pratiques plus courantes dans le cas d'expérimentations littéraires par exemple.

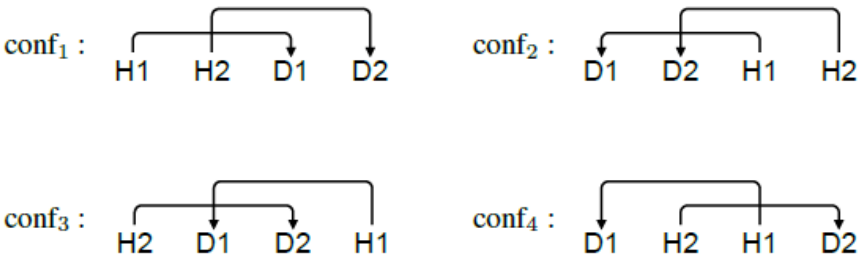


Fig. 80 — configurations de croisement de dépendances (Perrier, 2021 : 44)

Perrier (2021 : 49) calcule ainsi que la proportion d'arbres non projectifs dans la version UD de Sequoia n'atteint que 2,13 %. Dans notre cas, nous avons pu relever 12 arbres non projectifs[71], pour 219 phrases, donc 5,48 %.

Du point de vue linguistique, les dépendances non projectives peuvent être la conséquence de phénomènes syntaxiques complexes. Cependant la plus grande difficulté est technique dans le cadre d'un parser pour la prédiction de ces dépendances : selon Nivre et Nilsson (2005 : 100), il n'existe que peu de parsers capables de produire des structures de dépendance non projectives. La solution généralement adoptée — c'est le cas de notre parser[72] — pour permettre la prise en compte d'arbres non projectifs est alors une transformation complexe des arbres permettant de les rendre « pseudoprojectifs » (Nivre et Nilsson, 2005). Or, la précision de cette transformation n'est pas parfaite. Dans notre cas, le traitement a été particulièrement inefficace : aucune des 12 structures non projectives n'a été détectée, et le parser n'a pas réalisé une seule prédiction qui aurait rendu l'arbre syntaxique non projectif. Nous pouvons donc estimer que les erreurs sont au moins en partie dues à des difficultés techniques induites par des phénomènes linguistiques complexes.

[71] Il s'agit des arbres concernant les phrases PA-1, G-5, SA-5, RI-4, BT-9, TE-5, AA-5, NV-3, NV-6, NV-14, EF-5 et TPB-7.

[72] https://spacy.io/api/dependencyparser.

Du point de vue linguistique, nos structures peuvent être classées en deux grands groupes de structures pouvant entraîner des dépendances non projectives selon les règles d'annotation UD : les dépendances non projectives liées à un déplacement (ou détachement) d'un complément et celles liées aux structures *xcomp* pour lesquelles le complément d'objet direct est antéposé (voir infra).

a) Complément ou modificateur détaché

Ces dépendances non projectives liées à un déplacement sont majoritaires, en incluant 7 phrases sur 12. G-5 en constitue le prototype (fig. 81).

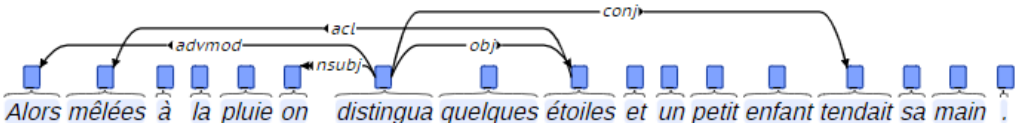


Fig. 81 – GE-5, prototype de non-projectivité liée à un déplacement

Nous pouvons constater ici deux couples de dépendances non projectives : « *alors* <advmod *distingua* » et « *mêlées* <acl *étoiles* » ainsi que « *mêlées* <acl *étoiles* » et « *distingua* conj> *tendait* ». L'élément commun ici est bien sûr la dépendance entre *mêlées* et *étoiles* : le premier, participe passé, placé devant le verbe *distingua*, a la fonction d'épithète détachée d'*étoiles*, complément d'objet direct de *distingua*. Nous pouvons nous assurer que le gouverneur de *mêlées* est effectivement *étoiles* en raison de son genre féminin et son nombre pluriel. Puisque les deux éléments liés sont placés de part et d'autre du verbe, toute relation de *distingua* dont le gouverneur (ici le verbe est *racine*, mais la dépendance virtuelle à la racine, considérée parfois comme un nœud vide à gauche de la phrase aurait été problématique également) ou le dépendant est placé à l'extérieur du couple *mêlées-étoiles* (c'est-à-dire à gauche de *mêlées* ou à droite de *étoiles*) croise la relation entre ces deux derniers, impliquant une non-projectivité. C'est donc le cas du complément adverbial *alors* et de la coordination des verbes.

Nous dénombrons **4** cas d'épithète détachée (PA-1, G-5, SA-5 et AA-5), **1** complément déterminatif détaché (BT-9), **1** apposition détachée (TE-5), et 1 répétition du complément d'objet proclitique *me* dont la seconde réalisation est rejetée derrière le verbe (NV-14). Il s'agit dans tous ces cas de déplacement par-delà le verbe constituant une structure de dépendance entourant celui-ci, avec un des éléments étant situé à sa gauche et l'autre à sa droite.

Les prédictions erronées des dépendances non projectives dans ce cas sont incohérentes et ne montrent pas de comportement systématique. À titre d'exemple, nous

fournissons au lecteur la prédiction de PA-1 (fig. 82), qui, face à la non-projectivité de la dépendance de l'épithète détachée *tel*¸ sélectionne cet adjectif comme racine de la phrase alors que celle-ci contient trois verbes conjugués à un mode non fini, ce qui en fait des candidats plus évidents.

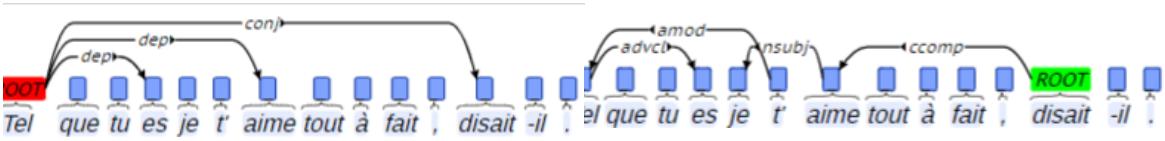


Fig. 82 — Prédiction et correction de PA-1

b) Complément d'objet direct antéposé dans une construction xcomp

Les structures *xcomp* dont le verbe à l'infinitif est transitif sont également susceptibles de présenter des dépendances non projectives. En effet, puisque dans cette situation UD prévoit que le sujet dépend du semi-auxiliaire et l'objet de l'infinitif, nous obtenons une structure prototypique à trois dépendances : « [*sujet*] <nsubj [*semi-auxiliaire*] », « [*semi-auxiliaire*] xcomp> [*infinitif*] » et « [*infinitif*] obj> [*complément d'objet direct*] ». Dans ce cas, lorsque l'objet est placé à gauche du sujet ou que le sujet est placé à droite de l'objet, des dépendances non projectives apparaissent.

Les deux situations dans lesquelles ces conditions sont réunies dans notre corpus sont dans des structures *xcomp* ayant subi une inversion du sujet et du verbe à cause d'une interrogation (RI-4 [fig. 83]) ainsi que dans des propositions relatives dans lesquelles le pronom relatif est complément d'objet direct (NV-3 [fig. 84], TPB-7 et EF-5).

Dans la première situation (fig. 83), le pronom interrogatif placé en début de proposition a le rôle de complément d'objet direct de l'infinitif et le sujet est rejeté après le verbe en raison de l'inversion sujet-verbe dans l'interrogation commençant par *que* en français (Grevisse et Goosse, 2016 : 542). Les dépendances *nsubj* et *obj* se croisent donc. La dépendance objet est également non projective avec tous les éléments dépendant du semi-auxiliaire (conjonctions de coordination, conjonctions de subordination, compléments de phrase…) placés à gauche de l'objet.

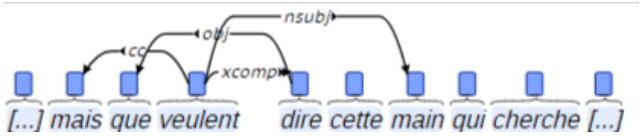


Fig. 83 — Non projectivité en fonction *xcomp* interrogative (RI-4)

Les croisements liés aux propositions relatives (fig. 84) suivent la même logique : la relation entre l'antécédent et le verbe de la proposition relative croise la dépendance du complément d'objet direct s'il est antéposé, ce qui est systématiquement le cas lorsque le pronom relatif est objet puisqu'il est toujours placé en tête de proposition. Nous voyons donc, pour NV-3, un croisement entre les dépendances « *main* acl:relcl> *voulu* » et « *que* <obj *reparaître* ». Puisque ces structures ne semblent pas particulièrement marquées ou expressives, nous pouvons estimer que l'apparition de la non-projectivité dans ces cas assez communs constitue une faiblesse du modèle *xcomp* d'UD.

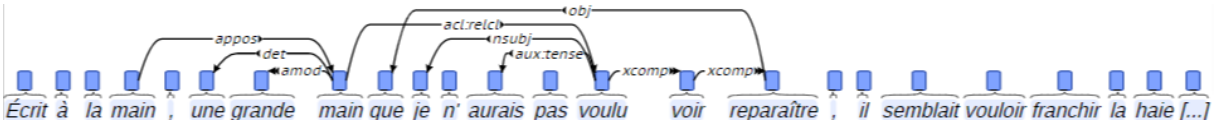


Fig. 84 — Croisement de dépendances dans une proposition relative (NV-3)

Il est intéressant de constater que le traitement est régulier dans cette sous-section : le pronom interrogatif ou le pronom relatif deviennent alors dépendants, avec l'étiquette sous-spécifiée *dep*, du semi-auxiliaire (fig. 85) ; cette stratégie n'a pas de valeur syntaxique autre que la conservation à tout prix de la projectivité.

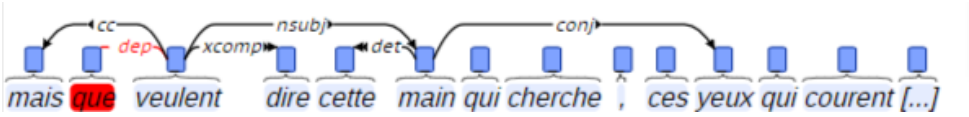


Fig. 85 — Prédiction de RI-4

4.2.14.2 Segmentation interne

Les erreurs de segmentation interne de la phrase ont déjà été introduites par l'intermédiaire de l'étiquette SEG, mais cette partie s'intéresse aux caractéristiques de ces erreurs.

	parataxis	*conj*	*appos*	*dislocated*
Nombre d'erreurs SEG par étiquette	8	11	4	1

Tab. 11 — Répartition des erreurs SEG par dépendance

Nous rencontrons 24 erreurs de ce type dans l'ensemble du corpus[73] ; il ne s'agit donc pas de cas isolés puisque cela représente un peu plus de 5 % du total. Il est intéressant de noter que ces erreurs n'apparaissent que dans quatre contextes, que nous

[73] C-2 (1), VV-6 (1), VV-9 (1), SA-6 (1), DC-3 (1), APN-3 (1), RI-4 (1), BT-2 (1), BT-8 (1), TE-6 (3), LNP-7 (1), TN-5 (4), HE-7 (5), NV-11 (1) et GL-1 (1).

avons définis comme *relations horizontales* : la parataxe, la coordination, l'apposition et la reprise syntaxique (tab. 11).

Le prototype de ces erreurs est celui de DC-4 (fig. 86) : une dépendance, *conj* (en cyan), n'est pas reconnue et le modèle prédit à la place deux arbres disjoints, chacun avec sa propre racine.

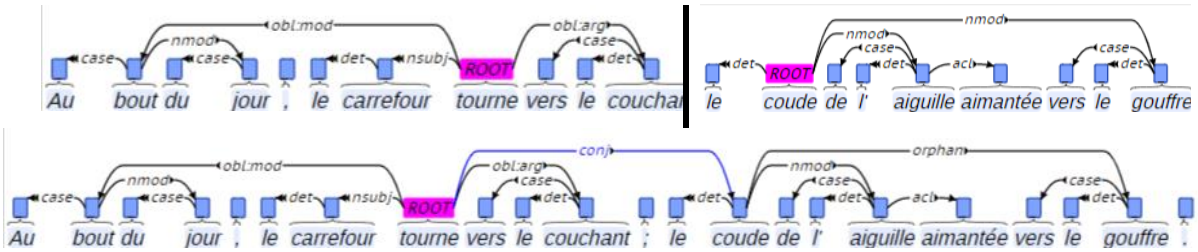


Fig. 86 — Prototype d'erreur de segmentation (DC-4)

L'étiquette *root*, accompagnée de la dépendance *0*, à la racine, n'est donc substituée qu'aux étiquettes *parataxis*, *conj*, *appos* et *dislocated*. Or, rappelons le fonctionnement de la prédiction : pour chaque étape de l'algorithme de transition, le FFNN branché à la fin des transformers assigne une valeur d'activation de chaque neurone de sortie, correspondant à chacune des possibilités de prédiction (le jeu d'étiquettes pour les étiquettes, les index des tokens de la phrase pour le choix du gouverneur). De toutes ces valeurs, qui ne sont que très exceptionnellement nulles, c'est celle qui est la plus élevée qui est sélectionnée comme prédiction. Cela signifie donc que, dans notre cas, lorsqu'une erreur SEG apparaît, c'est le neurone correspondant à la dépendance *0*, dépendance racine particulière, qui a la plus grande activation. La substitution de *root*[74] avec *parataxis* ou *conj* signifie donc que le token est *plus probablement* la racine qu'un élément de parataxe, coordination ou apposition. Cela peut se produire dans deux contextes : soit la prédiction *root* est particulièrement élevée parce que le token est convaincant à cet égard, soit la prédiction alternative est faible parce que le modèle est partagé entre différentes prédictions. Différents facteurs nous viennent à l'esprit lorsque nous considérons les raisons de la variation de l'équilibre entre la prédiction *root* et une prédiction alternative :

[74] Pour des raisons de clarté, nous mettons de côté la prédiction du gouverneur qui est pourtant probablement préalable à la prédiction de l'étiquette : *root* est en effet nécessairement associé à la dépendance vers l'index *0*, alors que l'index du gouverneur dans les autres cas est variable et cela complexifie le discours.

- L’augmentation de l’activation *root* due à la nature de l’élément : un verbe conjugué à un mode fini est très régulièrement racine de la phrase, de même que le substantif dans les phrases nominales. Dans le corpus d’entraînement, 63,6 % des racines sont des verbes, et 29,8 % sont des substantifs ou des noms propres. Dans notre corpus, 56,1 % des racines sont des verbes, et 31 % des substantifs. Nous constatons donc que les caractéristiques grammaticales et morphologiques peuvent entraîner une valeur élevée de la prédiction *root*. Il n’est donc pas surprenant que la totalité des erreurs de segmentation soit liée à des verbes (34,8 %) et des substantifs (65,2 %). Notons que les substantifs sont surreprésentés dans les erreurs de segmentation par rapport à l’entièreté du corpus.
- L’augmentation de l’activation *root* due à une division plus forte explicite marquée par un signe de ponctuation fort : nous constatons en effet que, des 8 points-virgules rencontrés dans le corpus, chacun précède une erreur de segmentation. De même, le tiret cadratin précède l’erreur dans 12 cas (pour 57 occurrences). Seules 2 erreurs de segmentation sont précédées de la virgule, et 1 n’est précédée d’aucun signe de ponctuation. La division autour du tiret cadratin est probablement renforcée par l’entraînement : il existe de nombreuses phrases dans le corpus d’entraînement qui sont issues de dialogues et débutent donc par cette marque de ponctuation.
- La diminution de l’activation de l’étiquette alternative due à l’accroissement de la taille de la dépendance (difficulté de la prédiction des dépendances à longue distance) : la moyenne de la longueur des dépendances brisées par ces erreurs de segmentation est supérieure à la moyenne du corpus (tab. 12).
- La diminution de l’activation de l’étiquette alternative due à une incertitude du modèle par rapport à cette étiquette : nous constatons en effet que le F-score est 0 pour *parataxis* et *dislocated*, 0,30 pour *appos* et 0,70 pour *conj* ; il s’agit donc d’étiquettes que le modèle reconnaît mal.

	Dépendances entre deux segments internes (erreurs SEG)	**Relations horizontales correctes du corpus**
Moyenne de la longueur des dépendances	12,04	7,57

Tab. 12 — MLD des relations horizontales

Une dernière condition nous semble nécessaire pour la prédiction d'une division de phrase : un flux de dépendance faible (taille de 1) à un point intermédiaire entre deux sous-arbres marqués. En effet, si la taille du flux est réduite à 1 entre deux sous-arbres avec une grande activité de flux, il n'est nécessaire de briser qu'une dépendance pour obtenir deux arbres syntaxiques disjoints, là où un flux de taille n > 1 nécessite au minimum n erreurs de choix du gouverneur pour disjoindre les sous-arbres. L'erreur de C-2 (fig. 87) est représentative d'une forte diminution du poids du flux entre deux sous-arbres avec un flux interne fort.

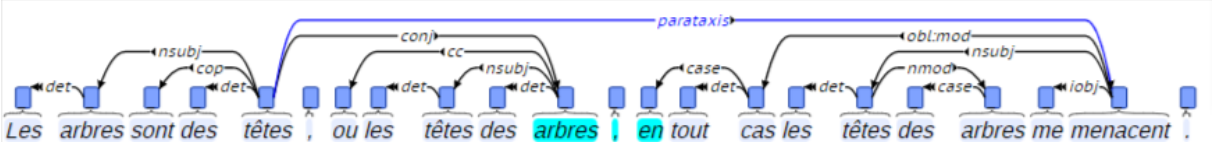


Fig. 87 — Correction (en bleu) de l'erreur de segmentation de C-2

En effet, en excluant la dépendance *parataxis* et la ponctuation, nous obtenons deux sous-arbres ayant un flux de dépendance de taille 2 en moyenne et de taille 4 au maximum. La dépendance *parataxis*, de longueur 14, bien au-dessus de la moyenne, est le seul lien entre les deux sous-arbres, avec un flux de taille 1 à la position (en cyan) entre *arbres* et *en* (ou entre la virgule et *en* si on prend en compte la ponctuation). Il semble donc s'agir d'une dépendance assez faible par sa longueur et par la faible taille (et poids) de flux en son centre.

Les sous-arbres sont moins marqués dans NV-11 (fig. 88) : le premier sous-arbre a un flux de dépendance de taille moyenne 1,16 et de taille maximale 2 ; le second sous-arbre a un flux de taille moyenne 1,61 et de taille maximale 4. De plus, la dépendance brisée, *parataxis* encore une fois, n'a une longueur que de 4 et semble donc plus forte. Cependant, nous voyons ici l'apparition du point-virgule qui semble particulièrement déterminant dans la prédiction avec 100 % de segmentation après ce signe de ponctuation.

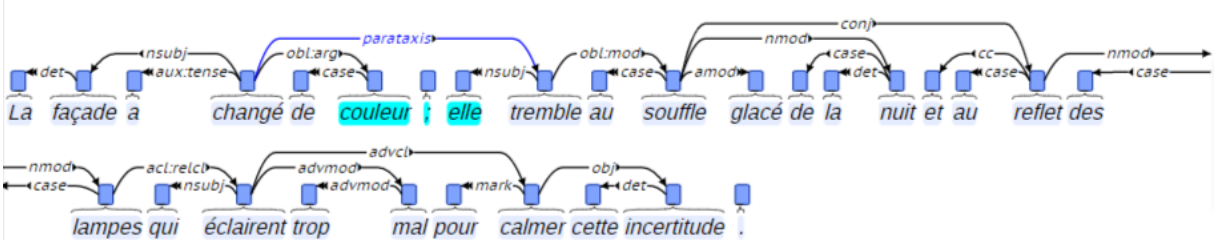


Fig. 88 — Correction de l'erreur de segmentation de NV-11

Il est toutefois important de garder à l'esprit que cette réduction de la taille du flux entre deux sous-arbres cohérents est très loin d'être une caractéristique exclusive des erreurs de segmentation : elle peut également apparaître dans tout cas de complément assez long, en particulier dans la plupart des subordinations puisque le verbe génère souvent une grande activité dépendancielle autour de lui en étant gouverneur des actants et des circonstants. Nous n'avons pourtant repéré aucun cas d'erreur SEG dans des cas de subordination : la présence d'une conjonction de subordination est vraisemblablement suffisamment importante pour que la prédiction *root* ne puisse se démarquer.

Du point de vue de la taille du flux, le cas de TN-5 (fig. 89) — qui est d'ailleurs la seule phrase dont les dépendances brisées sont des relations appositives — est interpellant : Dans la correction, la taille du flux à la première segmentation erronée est de 4, ce qui nécessite donc la rupture de 4 dépendances.

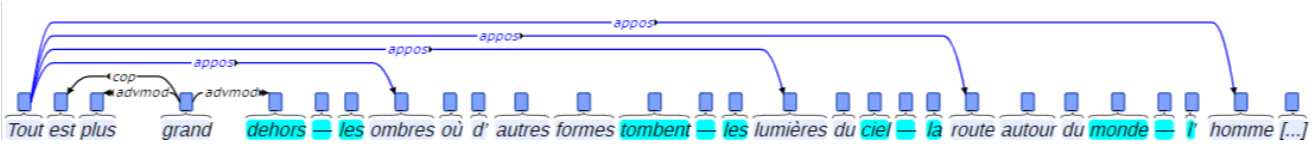


Fig. 89 — Erreurs de segmentation (en bleu) dans TN-5

Il est hautement improbable que le modèle, en considérant un bouquet d'appositions à quatre éléments, prédise la rupture des quatre dépendances et fasse des appositions des arbres disjoints. Nous soutenons donc comme hypothèse que la prédiction était partagée entre une segmentation, une apposition en bouquet et une apposition en chaîne (*ombres* dépendant de *tout*, *lumières* de *ombres*, *route* de *lumières* et *homme* de *route*), ce qui permet de retomber sur une structure avec une position intermédiaire à laquelle le flux est de taille 1 à chaque fois. La cohabitation des deux possibilités (puisque, contrairement à des adjectifs épithètes qui ne peuvent pas s'organiser en chaîne par exemple, une apposition, à noyau nominal, peut dépendre de tout nom, ce qui permet l'enchaînement) est peut-être un facteur ayant affaibli la prédiction de la dépendance apposition, ce qui a pu profiter pour que la prédiction *root* se démarque.

Quoi qu'il en soit, il est important de rappeler qu'aucun des facteurs n'est suffisant seul, et que même les considérations de flux sont à nuancer : le parser fonctionne sur un système de transition et prédit donc une dépendance à la fois. Le flux reste présent, même inconsciemment lorsque nous parlons, et il est certainement pris en compte dans la prédiction par la considération du contexte et de la phrase en un seul bloc par le

modèle, mais les dépendances qui sont prédites en dernier exercent toujours une influence limitée au moment de la prédiction des premières : il n'est donc pas certain que le modèle considère même la possibilité d'un flux de taille 4 pour TN-5. Ces réflexions de flux sont vraisemblablement plus importantes dans le cas des parsers basés sur la prédiction d'une traite du graphe entier (*graph-based parsers*, → 2.2).

4.3 Conclusion de l'évaluation linguistique

En guise de conclusion, puisque la discussion des résultats se trouve en réalité disséminée dans l'exposé, nous souhaitons résumer les faiblesses dégagées grâce à l'analyse linguistique. La présentation s'articule autour de quatre points : le modèle statistique, le corpus littéraire, les règles d'annotation UD et la langue en général. Bien sûr, ces catégories sont perméables : une faiblesse que nous attribuons plutôt au modèle peut également être influencée par UD, et les particularités du corpus littéraire sont *a fortiori* des particularités linguistiques également.

a) Faiblesses dues au modèle et son entraînement

Tout d'abord, nous remarquons qu'une quantité non négligeable d'erreurs pourraient être attribuées à un manque d'entraînement en raison de la rareté de certaines constructions, comme l'interrogation, le discours rapporté ou l'interjection, dans Sequoia-Train. Le corpus d'entraînement est donc largement améliorable.

Ensuite, la force de cette méthode est d'être basée sur des méthodes statistiques complexes permettant de s'adapter à des situations grâce à des similarités. Il en découle cependant plusieurs faiblesses : **1.** la similarité peut être trompeuse, comme cela est le cas entre la proposition relative (*acl:relcl*) et la structure des phrases clivées (*advcl:cleft*), ce qui peut entraîner des erreurs ; **2.** lorsque le modèle est face à une structure linguistique d'une grande originalité, la prédiction est très souvent extrêmement incohérente. Ce deuxième point est particulièrement bien illustré dans les situations dans lesquelles le modèle décompose analytiquement des figements n'ayant plus de cohérence syntaxique interne plutôt que les considérer d'un seul tenant.

Nous pouvons également constater que les difficultés techniques liées à la projectivité des phrases n'ont pu être contournées dans notre cas, malgré l'usage de transformations des dépendances non projectives.

Enfin, l'apparition d'erreurs comme celles de ponctuation nous rappelle que les outils utilisés sont des *boîtes noires*, que la conception de raisonnement ne s'applique

pas et que cela peut mener à des incohérences spectaculaires, sans que le modèle soit capable d'autocritique. Cela nous montre bien que, lorsqu'on parle de la phase d'*apprentissage* d'un modèle, il ne s'agit que d'une phase d'*ajustement des paramètres*.

b) Complexité particulière du corpus

En ce qui concerne le corpus, ses particularités majeures qui ont pu perturber la prédiction sont la présence d'une grande quantité de relations horizontales diverses, l'usage régulier d'ellipses verbales ainsi que le détachement régulier des compléments, allongeant donc les relations de dépendance dans la linéarité de la phrase.

c) UD, un environnement idéal pour l'analyse syntaxique automatique ?

À cette question, nous répondons par la négative bien que des analyses plus poussées doivent être menées pour confirmer ceci. En effet, tout d'abord, plusieurs incohérences, comme la concurrence entre les étiquettes *ccomp* (proposition complétive) et *parataxis* (parataxe) pour la description du discours rapporté sont problématiques pour la généralisation de l'analyse. De même, le fait que la description syntaxique des constructions à auxiliaire ou semi-auxiliaire (causatif, temps explétifs et fonctions *xcomp*) ne soit pas superposable conduit à des erreurs en chaîne ; nous estimons que l'analyse de SUD, dans laquelle l'auxiliaire est dans tous les cas la tête du syntagme verbal et seule l'étiquette change, est beaucoup plus pertinente pour diminuer le nombre d'erreurs.

Surtout, c'est le choix des têtes lexicales pour les syntagmes prépositionnels dans UD que nous critiquons : pour nous, cela nuit tout à fait à la reconnaissance des syntagmes prépositionnels et entraîne de nombreuses confusions entre *nmod* (complément déterminatif) et *appos* (apposition), qui sont pourtant des relations dont la différence est explicite sur le plan syntaxique. Ceci est également d'application pour la différence entre parataxe, coordination et subordination de propositions. Selon nos conclusions empiriques, nous estimons donc (sous réserve de vérification) que l'environnement SUD est plus pertinent pour l'annotation syntaxique, mais pourrait surtout être plus efficace pour l'analyse automatique.

d) Limites linguistiques du parsing

Enfin, nous avons également pu constater une série de limites linguistiques qu'il est impossible au parser de franchir. Il s'agit de situations qui sont complexes sur le plan

théorique, comme la distinction entre les phénomènes de coordination : aucune solution idéale de description en dépendance de la coordination n'a encore été proposée, et le fait que sa structure, dans les conceptions asymétriques, ne soit pas superposable avec la coexistence entraîne une grande quantité d'erreurs en cas de confusion.

L'évolution de la langue, les figements et leur démotivation syntaxique sont également une difficulté théorique : il s'agit d'objets linguistiques divers dont la considération analytique ou synthétique fait débat au sein de la communauté des linguistes. Sans un consensus, il est impossible d'établir une liste exhaustive des figements qui ne possèdent plus de structure syntaxique interne et qu'il faut considérer synthétiquement ; cette liste serait de toute façon rendue obsolète par le temps.

Pour finir, l'ambiguïté et l'homonymie sont d'autres universaux du langage pour lesquels une description syntaxique n'est pas toujours suffisante ni efficace.

Nous avons cependant pu constater que les erreurs liées à des limites linguistiques ne sont pas majoritaires, et que le chemin est encore long avant que le TAL n'atteigne cette frontière théorique.

5 Parsing syntaxique et littérature

Dans cette section, nous argumentons brièvement en faveur de l'intérêt de la littérature pour les études linguistiques (→ 5.1) et présentons les possibilités offertes par le parsing syntaxique pour les études littéraires, en particulier la stylistique littéraire (→ 5.2).

5.1 De l'importance de la littérature en linguistique

Pendant des siècles, la citation littéraire a occupé une place de choix dans les grammaires, aussi diverses soient-elles : quoi de plus satisfaisant, de plus valorisant pour un grammairien que d'illustrer un concept grammatical en s'appuyant sur les *grands auteurs* ? C'est ainsi qu'une grammaire de référence comme le *Bon Usage* (Grevisse et Goosse, 2016) intègre une énorme quantité d'extraits et citations littéraires comme exemples des constructions rencontrées en français. Il suffit, pour s'en convaincre, de consulter l'index des auteurs présenté à la fin de l'ouvrage ; de ce point de vue, c'est même la littérature qui devient l'objet d'étude. Marchello-Nizia et Petiot (1977) relèvent que, même dans la plupart des grammaires pédagogiques du XX^{e} siècle, les exemples sont très souvent littéraires. Seules trois grammaires des années 1970

destinées aux collégiens offrent des exemples plus variés, avec l'apparition de fragments de conversation, d'allusions à des émissions télévisées, et même d'une publicité pour un laxatif (Marchello-Nizia et Petiot, 1977 : 104). Il est donc indéniable que les grammairiens ont considéré les textes littéraires comme lieux du *bien écrire* et de la variété linguistique nécessaire à l'explication des phénomènes grammaticaux dans leur ensemble.

Au contraire, le linguiste, et la linguistique en général, a semblé s'éloigner de plus en plus de la littérature :

> D'autre part, nous l'avons vu, le linguiste se forge ses exemples, il n'a plus la position du « greffier » que, dans sa préface au Bon Usage, F. Desonay donne comme celle du grammairien face aux écrivains. (Marchello-Nizia et Petiot, 1977 : 107)

Il privilégie alors les exemples minimaux, avec le quasi-monopole des positions actancielles occupées par *Jean* et *Marie*. C'est ainsi également que les autres genres, qu'ils soient textuels ou non, ont été progressivement intégrés par la discipline comme matériau. Ne nous détrompons pas, c'est bien là une étape primordiale : qu'il s'agisse du genre informatif, du slogan, des messages vocaux ou de la conversation orale courante, nous sommes face à un usage de la langue qu'il est nécessaire de décrire. Cependant, nous déplorons que cette transition ait été réalisée au détriment de la littérature, qui est quasiment absente des treebanks, par exemple : après tout, la littérature n'est-elle pas également linguistique ? Est-ce être conservateur, ou *réac'*, de centrer une étude autour d'un texte littéraire ?

Le TAL en particulier semble s'être affranchi de la littérature : BERT est entraîné grâce à Wikipédia ou des journaux, les treebanks intègrent plutôt des phrases issues de la communication au grand public (Sequoia) ou de l'oralité (Rhapsodie [Lacheret-Dujour *et al.*, 2019])… Or, comme nous l'avons montré par l'illustration des difficultés d'analyse syntaxique automatique, le texte littéraire reste un lieu privilégié d'expérimentation, de variété et de rencontre de structures déroutantes qui sont néanmoins des usages de la langue. Nous plaidons donc sa cause ; peut-être serait-il bon d'accorder à nouveau une place au genre littéraire dans le traitement automatique des langues[75] ?

[75] Cette réflexion prend son point de départ dans une conversation informelle avec le Pr N. Mazziotta. Nous lui en donnons le crédit et le remercions.

Dans l'entreprise de la création d'un treebank littéraire, il ne faudrait pas sous-estimer alors l'importance de l'exégèse : courantes sont les structures ambiguës qu'il est difficile d'analyser syntaxiquement sans une compréhension solide du texte. BT-2 et les compléments *au tapis vert rayé* et *au rouge du couchant* en sont une illustration : sont-ce des compléments déterminatifs de *banquette, ressacs* ou *secousses,* ou encore des compléments adverbiaux du verbe passé sous ellipse ?

> (BT-2) Sur la banquette des ressacs et des secousses au tapis vert rayé, au rouge du couchant, le velours effacé et les jambes pendantes ; la fatigue de la mémoire endormie sous la mousse.

Un problème se pose cependant : ce type d'analyse participe à la désambiguïsation du texte, mais l'objectif du texte littéraire n'est-il pas souvent de construire une polysémie complexe ? Il s'agit là de questions fondamentales de la littérature qu'il faut prendre en compte.

5.2 L'apport du parsing syntaxique pour les études littéraires

La réciproque nous semble également vraie : le TAL peut apporter du soutien à l'étude des textes et des auteurs. Il ne s'agit évidemment pas d'une quelconque *exégèse automatique* qui, dans l'état actuel des connaissances, ne serait vraisemblablement d'aucune aide pour la compréhension des textes. Nous pensons cependant qu'une discipline en particulier, la stylistique, pourrait se nourrir des nouvelles méthodes automatiques. Molinié définit l'objectif de la stylistique littéraire :

> Inutile de faire semblant de ne pas savoir ce qu'on cherche : caractériser une manière littéraire à la différence d'une autre, qu'il s'agisse de différences d'auteurs, d'œuvres ou de genres. On pose le postulat suivant : une manière littéraire est le résultat d'une structure langagière. Décrire une structure langagière, c'est démonter les éléments qui la composent, mais auxquels elle ne se réduit pas, et mettre au jour les diverses grilles qui organisent ces éléments. […] D'autre part, on ne considère que des procédés, des moyens d'expression, des déterminations strictement formelles. Mais aussi jouant au niveau de la forme de l'expression, le [*sic*.] stylistique touche forcément la forme du contenu. La pratique stylistique ne peut donc être que structurale. (Molinié, 1986 : 12)

L'auteur avait déjà également senti l'intérêt de l'approche quantitative appliquée aux phrases :

> Une des méthodes de l'analyse de discours est la quantification. […] le champ d'études a été le mot, et c'est normal ; on a donc fait beaucoup de lexicologie quantitative […]. On est en train de déborder la lexicologie quantitative par l'étude

> des segments répétés. Il était temps, en effet, de sortir de l'unité-mot pour envisager des groupements séquentiels, même non syntagmatiques […]. C'est une ouverture réellement novatrice, qui portera sans doute des lumières inattendues sur divers textes.
>
> Mais, à notre avis, il faut aller encore plus loin. L'horizon des recherches appliquées aux segments répétés doit être l'étude de la phrase. (Molinié, 1986 : 184-185)

C'est vers là que nous souhaitons emmener le lecteur : le texte littéraire, au même titre que toute réalisation linguistique, est nécessairement soumis à une organisation syntaxique. Or, l'écart par rapport à ce qu'on pourrait appeler un *degré zéro* de la syntaxe est, pour nous, un élément constitutif et définitoire du *style* entendu par la stylistique. Une analyse des structures syntaxiques récurrentes est alors particulièrement intéressante pour la description d'un genre, d'un auteur ou d'une œuvre **relativement** à un autre texte. Cependant, ceci ne peut se faire qu'à une condition : la disponibilité d'une grande quantité de données, autant en ce qui concerne l'objet du travail que toutes les autres productions langagières. C'est ici que nous pouvons tirer profit de la machine et du TAL : des parsers efficaces permettent d'accélérer considérablement le processus d'annotation syntaxique des corpus et des méthodes quantitatives peuvent ensuite prendre le relais.

Une fois les données disponibles, il est possible de réaliser une quantité incroyable d'analyses permettant d'accroître la connaissance des textes : d'une typologie des compléments de phrase à une analyse de la ponctuation[76], les possibilités ne sont limitées que par l'imagination des stylisticiens. Nous pouvons même, en dépendance, prendre en considération le flux de dépendance. À titre d'exemple, nous aimerions proposer une analyse très brève du squelette de la phrase et de ses propositions chez Reverdy (→ 5.2.1), ainsi que dégager et décrire une série de ruptures qui peuvent être ressenties lors de la lecture (→ 5.2.2). Puisqu'aucun autre corpus littéraire en français annoté selon UD n'est à notre disposition, la comparaison ne peut-être faite qu'avec Sequoia. Cela enlève de la valeur à la comparaison et nous aurions préféré comparer

[76] En dépendance, nous pourrions imaginer une analyse comme celle-ci : pour chaque marque de ponctuation, quelle est la (ou les) dépendance la plus courte (et donc généralement celle qui conditionne l'apparition de la ponctuation) reliant un élément à droite à un élément à gauche de la marque ? Ce type d'analyse serait particulièrement intéressante à mener chez Reverdy : le tiret cadratin est parfois présent entre des propositions juxtaposées, mais aussi parfois, d'une façon plus spectaculaire, entre le sujet et le verbe (voir HE-7) ou le verbe et ses compléments adverbiaux.

deux auteurs, ou deux recueils de Reverdy, mais ces éléments sont suffisants à titre d'illustration.

5.2.1 Le squelette de la phrase dans *Flaques de verre*

Le premier paramètre qu'il nous semblait intéressant de présenter est celui du *squelette* de la phrase. Nous entendons par là « quelle est la racine de la phrase ? » et « comment la phrase se construit-elle autour de cette racine ? ». Lorsque le corpus est, comme ici, annoté en dépendance, cela revient à considérer quelles sont les relations des dépendants du token *root*, et quelle est la nature de celui-ci. Notre objectif a donc été d'isoler ces relations. Nous ne nous sommes cependant pas limité à cela : puisqu'il arrive qu'un autre nœud soit rattaché à la racine selon une relation d'équivalence ou d'entassement, nous avons décidé d'inclure les *pseudoracines*, c'est-à-dire le nœud principal de toute proposition ou syntagme en relation horizontale avec la racine. Puisqu'il peut s'agir autant de syntagmes verbaux que nominaux, nous choisissons *syntagmes équivalents* pour désigner autant le syntagme de la racine que celui des pseudoracines. Par exemple, dans *Pierre mange et Marie boit*, nous considérons tout autant les racines et pseudoracines *mange* et *boit* comme appartenant au squelette de la phrase, ce qui définit les syntagmes équivalents *Pierre mange* et *Marie boit*. Il s'agit dans ce cas de deux nœuds verbaux. Le squelette de cette phrase est alors celui-ci : [nsubj->VERB]---[nsubj->VERB][77].

En plus d'effectuer cette analyse sur les phrases, nous avons isolé les différents types de syntagmes équivalents afin de considérer leur fréquence. Ces catégories sont basées sur l'égalité stricte plutôt qu'un ratio de similarité, ce qu'il serait intéressant de développer. Nous avons rapidement rencontré des difficultés en raison de la quantité de particules entourant la racine : par exemple, la particule explétive *ne* de BT-4 (*Rien ne sort*) ajoute une étiquette (*expl*) au squelette de la phrase sans être déterminante pour notre analyse. Nous avons donc décidé dans un second temps de réduire la diversité des analyses : nous avons rassemblé tous les compléments adverbiaux non essentiels (propositions [*advcl*], syntagmes prépositionnels [*obl:mod*] et adverbes [*advmod*]) sous

[77] Cette notation nous est propre. Les crochets (**[**nsubj->VERB**]**) indiquent une proposition ou un syntagme équivalent. La forme en majuscules ([nsubj->**VERB**]) correspond à la nature de la racine ou pseudoracine. Les flèches (-> et <-) indiquent le sens de la dépendance, c'est-à-dire si l'élément qui reçoit cette relation est placé avant ou après la racine dans la linéarité de la phrase ; elles sont optionnelles. Les trois tirets (---), optionnels également, divisent visuellement les paires de crochets.

l'étiquette *advmod* ainsi qu'ignoré les particules explétives, les auxiliaires de temps[78], les déterminants numéraux, et enfin les prépositions et conjonctions de subordination apparaissant dans une proposition principale.

Le résultat de l'analyse des phrases entières est un peu décevant : en raison de la diversité possible, et en l'absence d'un critère de *similarité suffisante* remplaçant le critère d'égalité parfaite, la plupart des squelettes sont uniques ; nous ne les présentons donc pas ici. L'ensemble des résultats est à trouver en annexes (M.3). L'analyse des syntagmes équivalents réduits produit, elle, des résultats plutôt intéressants : entre les corpus, nous constatons une grande différence de fréquence des dix squelettes les plus représentés chez le poète (Tab. 3).

	Reverdy	**Sequoia**
[nsubj->VERB]	6,27 %	0,10 %
[nsubj->VERB<-advmod]	4,96 %	0,45 %
[det->NOUN]	4,44 %	0,54 %
[det->NOUN<-nmod]	4,18 %	0,89 %
[nsubj->VERB<-obj]	3,39 %	3,08 %
[nsubj->VERB<-obl:arg]	3,13 %	1,14 %
[advmod->nsubj->VERB<-obj]	2,35 %	1,56 %
[det->NOUN<-acl]	2,35 %	0,13 %
[det->NOUN<-acl:relcl]	2,09 %	0,13 %
[PRON]	2,09 %	0 %

Tab. 13 — Comparaison avec Sequoia de la fréquence des squelettes les plus représentés dans Reverdy

Tout d'abord, nous constatons que les squelettes les plus fréquents chez Reverdy sont aussi les plus simples. Cela est particulièrement surprenant lorsque nous constatons la fréquence de 6,27 % du squelette [nsubj->VERB], contre seulement 0,10 % dans Sequoia. En effet, ce type de proposition tout à fait minimale contenant simplement un sujet et son verbe semble très peu courant dans le genre de Sequoia, contrairement aux textes du poète. Nous pourrions émettre deux hypothèses concernant un usage littéraire potentiel, qu'il serait nécessaire de confirmer grâce à un plus grand nombre de données : **1.** la présence de propositions et syntagmes équivalents aussi courts et directs que ceux-ci participe à la variation de rythme souvent appréciée par les auteurs ; **2.** l'absence même du complément d'objet direct participe à un fonctionnement de distillation au compte-gouttes des compléments permettant de tenir le lecteur dans un état d'attente,

[78] *J'ai couru* a ainsi le même squelette que *je cours*. Nous n'avons cependant pas ignoré les auxiliaires de voix passive (*je suis mangé* est différent de *j'ai mangé*).

d'interrogation et d'interprétation de l'implicite. Du moins, l'auteur considère qu'il *n'est pas nécessaire* d'utiliser des prédicats complexes, ce qui met l'accent plutôt sur le sujet. Ceci pourrait être caractéristique de Reverdy, qui omet régulièrement le verbe, éliminant donc tout à fait le prédicat dans ce que nous avons appelé la description ontologique (→ 4.2.2 et 5.2.2). Les exemples ci-dessous présentent ces squelettes autant en tant que phrase complète qu'en tant que syntagme équivalent.

(RI-2) **Les feuilles de soleil se détachent** et **volent — les carrés d'or se posent** et **les reflets des glaces se décollent**.
(LPN-7) Les plus noirs, les plus mous, les plus vagues sont venus de plus loin — mais… **les instruments de la fanfare éclatent**, **le chef d'orchestre tombe**, les fenêtres qui s'épanouissent et les fleurs se noient dans un nouveau silence, car ici il n'y a pas d'autre air.
(PA-4), Mais **le soleil persiste**.
(AR-2) **Il pleure**.
(VV-2) **L'atmosphère tinte**.
(CS-6) **Je crois**.
(BT-4) **Rien ne sort**.
(LPN-8) **Le morceau continue**.

Ensuite, la fréquence des phrases (ou du moins, des syntagmes équivalents) nominales est particulièrement élevée : rien qu'en considérant ces dix squelettes, 15,15 % des syntagmes équivalents sont nominaux ou pronominaux. FP-6 et VV-1 sont des exemples contenant à plusieurs reprises des squelettes [det->NOUN] et [det->NOUN->nmod]. En l'absence d'un prédicat, il n'y a encore une fois pas de prédication.

(FP-6) Les yeux et la parole.
(VV-1) Le grelot de la lune, la pointe du kiosque et la boule du toit.

Enfin, nous pouvons noter que seules les propositions minimales dont le complément adverbial est postposé ([nsubj->VERB<**-advmod**], voir G-4) sont fréquentes dans le corpus, et largement plus que dans Sequoia. En effet, seule une occurrence (0,26 %) du squelette [advmod->nsubj->VERB] (voir NV-8) a été détectée, pour 3 (0,10 %) dans Sequoia. La tendance s'inverse lorsque le verbe possède un complément d'objet direct : le complément adverbial est antéposé ([advmod->nsubj->VERB<-obj], voir VF-7) dans 2,35 % des cas (1,56 % dans Sequoia), et postposé (voir DC-4) dans 1,57 % des cas (0,95 % dans Sequoia).

(G-4) **Nous étions dehors** et il pleuvait
(NV-8) Une porte s'ouvrait, **au fond de l'avenue le ciel se détendait** et le vent, qui venait de l'autre côté du sol, faisait flotter les franges des tentures.
(VF-7) **Cependant quelques visages idiots conservent leur sérieux** car, pour eux, rien ne compte, mais eux.
(DC-4) Sous la voûte, les boules qui sortent du clocher roulent dans tous les coins et **les heures tracent une lumière au milieu des arcs et des moulures**.

C'est maintenant, comme le dit Muller (1979 : 227-229), au statisticien de prendre le relais afin de confirmer que ces modèles ne sont pas (ou du moins, ne sont *probablement* pas) le fruit du hasard.

Le corpus analysé n'est, pour nous, pas suffisamment étendu pour fournir des conclusions solides sur le style du recueil, de même que Sequoia est loin d'être une référence intéressante de comparaison. Nous ne poussons donc pas l'analyse des fréquences plus loin, mais nous espérons avoir partagé au lecteur ce sentiment qui est nôtre que ce type de comparaison est tout à fait pertinent pour le développement d'une stylistique syntaxique. Passons maintenant à l'analyse des ruptures.

5.2.2 Ruptures d'équivalence et ellipses

Nous pouvons réutiliser les analyses précédentes pour proposer la notion de *rupture d'équivalence* et les repérer rapidement dans un texte : il s'agit de cas dans lesquels deux syntagmes sont en position d'équivalence alors que leur pseudoracine n'est pas de même nature. C'est par exemple le cas de coordinations entre un syntagme verbal et un syntagme nominal. Nous avons cependant écarté les situations dans lesquelles la pseudoracine est un adjectif ou un nom en raison d'une construction attributive (puisque la pseudoracine est alors l'attribut du sujet plutôt que la copule) : rien d'original n'est à signaler dans une coordination de type *Jean est malade et Marie va au marché*. Nous dénombrons 12 ruptures[79], dont trois exemples sont donnés ci-dessous (AR-7, PT-4 et SA-6) ; les pseudoracines de nature différente y sont en gras.

(AR-7) La face grimaçante se **détourne**, et, de l'autre côté de l'eau, une **forme** très blanche entre les arbres verts qui bougent.

(PT-4) Tout entière, la région **tourne** au fil du cadran — les **élans** des rayons dorés sous la paupière doublés par le bruit sourd de la vie des rivières et des

[79] AR-7, PT-4, VF-2, SA-6, DC-3, BT-8, TN-3, HE-6, HE-7, AP-8, GL-1 et GL-2.

pentes gardées par des plis remuants — jusqu'aux franges du ciel où fument les prières, dans la campagne grise et les cris du couchant.

(SA-6) Les éclairs sont **restés** debout sur le fond sombre — les **têtes** remuées en rond près des rideaux et les **visages** éclairés contre la vitre — les **yeux** ouverts qui n'ont jamais fini de regarder.

Puisque la plupart de ces constructions nous semblent être explicables par des ellipses, il nous reste à ajouter toutes les phrases dans lesquelles la relation *orphan* (qui, pour rappel, est la relation des éléments dépendant d'un élément effacé lorsque leur relation originale serait incohérente ou trompeuse avec la tête promue dans l'analyse) est rencontrée pour dénombrer tous ces cas. Le tableau ci-dessous (tab. 14) en reprend la liste organisée par pièce.

Poème	**Phrase(s)**	**Poème**	**Phrase(s)**	**Poème**	**Phrase(s)**
AR	7	VF	1, 2, 5, 9	AA	4
PT	1, 3, 4	SA	4,6	HE	1, 2, 6, 7
FP	1,3	DC	3	AP	8,13
G	7,8	BT	2, 8, 11	GL	1, 2, 3
CS	1,2	TN	3	TPB	15

Tab. 14 — Répartition des phrases à rupture ou ellipse

Ce type de phrase est présent dans 15 des 25 poèmes que nous avons analysés, ce qui en fait un élément récurrent, mais pas systématique. Notons que la densité (c'est-à-dire le nombre d'occurrences par rapport au nombre de phrases du poème) est variable : seule 1 phrase sur 9 du poème *D'une autre rive* (AR) satisfait à nos critères ; au contraire, le phénomène est très dense dans *L'homme aux étoiles* (HE, ci-dessous), puisque 4 des 7 phrases montrent des ruptures. Notons que dans deux tiers des cas, plus d'une phrase de ce type est présente.

(HE) **Une lampe dans chaque main**. **D'un bout de la chaîne aux étoiles**. Les fenêtres bleues du matin, le toit verni et l'escalier qui descend plus bas que la toile. Car il y a la mer entre le mur et l'homme et la nuit dépliée qui arrête le bruit. Il y a le bateau blanc qui écarte les lames et l'aile du soleil qui partage le vent.
Mais, surtout, le front troué par les épines, le cœur d'où sort la flamme et les yeux éplorés — le regard frappe au ciel et la porte qui s'ouvre laisse entrevoir l'espace où remuent les formes mortes sur les chemins tracés par un doigt lumineux.
Les arbres du jardin fermé sont sur la grille — les pointes du signal à côté de la mer — les deux battants ouverts sur l'horizon qui grince — le jour lâché — s'évade et piétine les ombres — les hommes — les étoiles tombées sur le revers.

Grâce au corpus annoté, nous avons pu dégager très rapidement cet ensemble de phrases particulières qui sont assez récurrentes dans le recueil *Flaques de verre* et observer leur répartition. Il a également été possible d'expliquer syntaxiquement ce qui est déroutant : l'auteur met en relation horizontale deux syntagmes n'ayant pas une pseudoracine de même nature ou omet le verbe faisant office de racine ou pseudoracine.

Notons qu'UD n'est pas l'environnement idéal pour effectuer ce type d'analyse stylistique. En effet, par exemple, lorsque nous nous intéressons à la linéarité de la phrase, à sa forme finale, le niveau syntaxique le plus intéressant est celui de la syntaxe de surface (Kahane, 2003). Or, des choix comme celui des têtes lexicales relèvent plutôt de la syntaxe profonde. Peut-être qu'un type d'annotation comme SUD, qui se veut explicitement comme une description de surface (***Surface-syntactic*** *Universal Dependencies*), serait plus adapté.

6 Conclusion

Dans ce travail, nous avons tout d'abord montré l'intérêt, et l'importance, de l'évaluation linguistique des modèles statistiques de TAL. Grâce à celle-ci, la critique peut ainsi dépasser l'entraînement ou l'architecture du modèle et nous permet de nous interroger autant sur le corpus d'entraînement, que l'environnement d'annotation ou les limites linguistiques inhérentes à l'acte même de parole. De cette façon, nous avons pu dégager un faisceau d'erreurs indiquant que UD, malgré sa popularité dans le domaine, ne semble pas constituer un cadre théorique idéal pour l'analyse syntaxique et *a fortiori* l'analyse syntaxique automatique.

Nous avons également défendu une position selon laquelle les études littéraires, et la littérature, sont importantes et pertinentes pour le développement du traitement automatique des langues. Le contraire nous semble également vrai : le TAL est susceptible de fournir des outils extrêmement puissants au service de la stylistique ou de la description des genres textuels et littéraires.

La limite la plus claire de cette étude est celle de son ampleur : en raison du temps nécessaire à une évaluation manuelle, l'analyse a été menée sur un corpus réduit. Il serait évidemment profitable d'étoffer celui-ci, autant en prenant en compte d'autres pièces de *Flaques de verre* qu'en intégrant d'autres recueils, d'autres auteurs, voire

d’autres genres. Également, toute analyse manuelle comporte une part d’arbitraire ; nous n’y échappons pas, notamment autour de la distinction entre cohérence et ambiguïté.

De plus, en raison de la multitude de facteurs impliqués lors de chaque prédiction, il est impossible de comprendre tout à fait le comportement du modèle. Ce travail ne permet donc que de dégager des hypothèses, des pistes de réflexion, qui enrichissent la réflexion autour des outils du traitement automatique des langues naturelles.

Enfin, il serait particulièrement pertinent de continuer ce travail en incluant une comparaison d’environnements d’annotation, comme UD et son concurrent SUD, ou encore en développant les liens que nous avons entraperçus entre le flux de dépendance, la complexité syntaxique et les erreurs de prédiction.

Nous finirons notre exposé sur une citation librement inspirée d’un entretien radiophonique d’Alan Turing à la BBC en 1951 :

> *The attempt to make a* [*speaking*] *machine will help us greatly in finding out how we* [*speak*] *ourselves*.

7 Bibliographie

Corpus

Reverdy, P. (1984 [1929]). *Flaques de verre*, Garnier-Flammarion.

Sources secondaires

Amidi A., Amidi S. (2018). *VIP Cheatsheet: Recurrent Neural Networks*, Stanford University.

Bretto, A., Faisant, A. et Hennecart, F. (2012). *Éléments de théorie des graphes*, Springer.

Candito M et Seddah D. (2012). « 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* [Juin 2012, Grenoble, France].

Charniak, E. (2018). *Introduction to deep learning*, The MIT Press.

Dalrymple, M. (2001). *Lexical Functional Grammar. Syntax and Semantics, volume 34*, Emerald Group Publishing Limited.

De Marneffe, M. C., et al. (2021). « Universal Dependencies ». *Computational Linguistics* : 251-308.

Dekang, L. (2003). « Dependency-Based Evaluation of Minipar ». Abeillé, A. (éd.). *Treebanks: Building and Using Parsed Corpora* : 317-329.

Devlin, J. et al. (2019). « Pre-training of Deep Bidirectional Transformers for Language Understanding ». *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (1) : 4171-4186.

Dozat, T. et Manning, C. (2017). « Deep biaffine attention for neural dependency parsing ». *5th International Conference on Learning Représentations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Feuillard-Aymard, C. (1989). *La Syntaxe fonctionnelle dans le cadre des théories linguistiques contemporaines*, Université Paris V.

Gerdes, K. *et al.* (2018). « SUD of Surface-Syntactic Universal Dependencies : An annotation scheme near-isomorphic to UD ». *Universal Dependencies Workshop 2018*.

Gerdes, K. *et al.* (2019). « Improving Surface-syntactic Universal Dependencies (SUD) : surface-syntactic relations and deep syntactic features ». *TLT 2019 – 18th internation Workshop on Treebanks and Linguistic Theories, Aug 2019, Paris, France*.

Grevisse et Goosse. (2016 [1936]). *Le Bon Usage*, De Boeck Supérieur.

Gómez-Rodríguez, C. (2017). « On the relation between dependency distance, crossing dependencies, and parsing ». *Physics of Life Reviews 21* : 200-203.

Gómez-Rodríguez, C. *et al.* (2018). « Global Transition-based Non-projective Dependency Parsing ». *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)* : 2667-2675.

Hadermann, P. (1993). *Étude morphosyntaxique du mot* où, Duculot.

Imrényi, A. et Mazziotta, N. (2020). « Aspects of the theory and history of dependency grammar ». *Chapters of Dependency Grammar. A historical survey from Antiquity to Tesnière*, John Benjanmins Publishing Company : 1-22.

Jiang, J. et Liu, H. (2015). « The effects of sentence length on dependency distance, dependency direction and the implications–Bases on a parallel English-Chinese dependency treebank ». *Language Sciences 50* : 93–104.

Kahane, S. (1997). « Bubble trees and syntactic représentations ». Becker et Krieger (éds). *Proc. 5th M. of Mathematics of language*, Saarbrücken.

Kahane, S. (2001). « Grammaire de dépendance formelles et théorie Sens-Texte ». *TALN 2001*, Tours.

Kahane, S. (2003). « The Meaning-Text Theory ». *Dependency and Valency, Handbooks of Linguistics and Communication Sciences*, De Gruyter.

Kahane, S. Yan, C. et Botalla, M.-A. (2017). « What are the limitations on the flux of syntactic dependencies ? Evidence from UD treebanks ». *4th international conference on Dependency Linguistics (Depling)*, Sep 2017, Pise, Italie : 73-82.

Kovář, V., Jakubíček, M., & Horák, A. (2016). On Evaluation of Natural Language Processing Tasks - Is Gold Standard Evaluation Methodology a Good Solution ?

Proceedings of the 8th International Conference on Agents and Artificial Intelligence : 540-545.

Lacheret-Dujour, A., Kahane, S. et Pietrandrea P. (éds). (2019). *Rhapsodie. A prosodic and syntactic treebank for spoken French*, John Benjamins Publishing Company.

Liu, H. (2007). « Probability distribution of dependency distance ». *Glottometrics* 15 : 1-12.

Liu, H. (2008). « Dependency Distance as a Metric of Language Compréhension Difficulty ». *Journal of Cognitive Science 9*(2) : 159-191.

Liu, Y. et al. (2019). *RoBERTa: A robustly optimized BERT pre-training approach.*

Marchello-Nizia, C. et Petiot, G. (1977). « Les exemples dans le discours grammatical ». *Langages 45* (mars 1977) : 84-111.

Martin, L. et al. (2020). « CamemBERT : a Tasty French Language Model ». *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, *Association for Computation Linguistics* : 7203-7219.

Mazziotta, N et Kahane, S. (à paraître). *L'émergence de la syntaxe structurale de Lucien Tesnière*.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*, State University Press of New York.

Merleau-Ponty, M. (2005). *Phenomenology of Perception*, Routlege.

Molinié, G. (1986). *Éléments de stylistique française*, PUF.

Muller, C. (1979). *Langue française et linguistique quantitative. Recueil d'articles*, Slatkine.

Nivre, J. et Nilsson, J. (2005). « Pseudo-Projective Dependency Parsing ». *Proceedings of the 43rd Annual Meeting of the ACL*, Association for Computational Linguistics : 99-106.

Nivre, J. et al. (2020). « Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection ». *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* : 4034-4043.

Osborne, T. et Gerdes, K. (2019). « The status of function words in dependency grammar: A critique of Universal Dependencies (UD) ». *Glossa: a journal of general linguistics 4*(1) : 17. 1-28.

Perrier, G. (2021). « Étude des dépendances syntaxiques non projectives en français ». *Revue TAL, ATALA (Association pour le Traitement Automatique des Langues) 62*(1) : 39-63.

Polguère, A et Mel'čuk, I. (éds). (2009). *Dependency in Linguistic Description*, John Benjamins Publishing Company.

Resnik, P. et Lin, J. (2010). « Evaluation of NLP Systems ». Clark, A. Fox, C. et Lappin, S. (éds). *The Handbook of Computational Linguistics and Natural Language Processing*, Wiley-Blackwell : 270-295.

Rohrmanstorfer, S. et al. (2021). "Image Classification for the Automatic Feature Extraction in Human Worn Fashion Data". *Mathematics 9* : 624.

Rosenblatt, F. (1958). « The Perceptron: a probabilistic model for information storage and organization in the brain ». *Psychological Review 35*(6) : 386-408.

Rossi-Gensane, N. (2017). « Syntaxe et paradigme(s) : outre les relations de dépendance, les relations d'équivalence », *Signata 8* : 65-99.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). « Multiword Expressions : A Pain in the Neck for NLP ». *Computational Linguistics and Intelligent Text Processing*, 1-15.

Sokolova, M. et Lapalme, G. (2009). « A systematic analysis of performance measures for classification tasks ». *Information processing and management 45* : 427-437.

Staudemeyer, R. et Morris, E. (2019). *Understanding LSTM, a tutorial into Long Short-Term Memory Recurrent Neural Networks*.

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

Tsarfaty, R. *et al.* (2011). « Evaluating Dependency Parsing: Robust and Heuristics-Free Cross-Annotation Evaluation ». *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* : 385-396.

Turing A. M. (1950). « Computing Machinery and Intelligence ». *Mind 49* : 433-460.

Vaswani A. et al. (2017). « Attention is All you Need ». *Advances in Neural Information Processing Systems 30 (NIPS 2017)* : 5998-6008.

Wolf, T. et al. (2020). « Transformers: State-of-the-Art Natural Language Processing ». *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, *Association for Computational Linguistics* : 38-45.

Yan, C. et Kahane, S. (2018). « Syntactic complexity combining dependency combining dependency length and dependency flux weight ». *Proceedings of the First Shared Task on Measuring Language Complexity* : 38-48.

Sitographie

« annodoc », *Github*, consulté en avril 2022. URL : https://github.com/spyysalo/annodoc

« tesseract », *Github*, consulté en septembre 2021. URL : https://github.com/tesseract-ocr/tesseract

« DependencyParser », *spaCy*, consulté en juin 2021. URL : https://spacy.io/api/dependencyparser

« fr_dep_news_trf », *spaCy*, consulté en mars 202. URL : https://spacy.io/models/fr#fr_dep_news_trf

« spaCy 101: Everything you need to know », *spaCy*, consulté en juin 2021. URL : https://spacy.io/usage/spacy-101

« advcl:cleft: cleft adverbial clause modifier », *Universal Dependencies*, consulté en avril 2022. URL : https://universaldependencies.org/fr/dep/advcl-cleft.html

« Ellipsis », *Universal Dependencies*, consulté en février 2022. URL : https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis

« fixed: fixed multiword experssion », *Universal Dependencies*, consulté en février 2022. URL : https://universaldependencies.org/u/dep/fixed.html

« parataxis: parataxis », *Universal Dependencies*, consulté en février 2022. URL : https://universaldependencies.org/u/dep/parataxis.html

« parataxis: Reported Speech », *Universal Dependencies*, consulté en avril 2022. URL : https://universaldependencies.org/u/dep/parataxis.html#reported-speech

« punct: punctuation », *Universal Dependencies*, consulté en février 2022. URL : https://universaldependencies.org/u/dep/punct.html

« Reported Speech », *Universal Dependencies*, consulté en avril 2022. URL : https://universaldependencies.org/u/dep/ccomp.html#reported-speech

« Universal Dependency Relations », *Universal Dependencies*, consulté en octobre 2021. URL : https://universaldependencies.org/u/dep/index.html

« Working Group on Comparative Constructions », *Universal Dependencies*, consulté en avril 2022. URL : https://universaldependencies.org/workgroups/comparatives.html

8 Annexes

Pour des raisons d'économie de papier et de cohérence par rapport au format des fichiers (.xlsx, .conllu…), les annexes qui ne sont pas essentielles à la lecture du travail sont disponibles uniquement sur Matheo. Il s'agit de la différence entre les annexes *A* et *M* : les premières sont comprises dans ce document alors que les secondes sont accessibles via le portail uniquement.

Liste des annexes A :

- A.1 : Corpus ;
- A.2 : Abréviations des titres des poèmes ;
- A.3 : Corpus divisé en phrases ;
- A.4 : Liste et exemples des étiquettes de relations de UD-Sequoia
- A.5 : Statistiques de comparaison des corpus.

Liste des annexes M :

- M.1 : Visualisations des arbres (M.1 Visualisations des arbres.pdf) ;
- M.2 : Paires d'étiquettes confondues (M.2 Paires.txt) ;
- M.3 : Squelettes de la section 5 (M.3 Squelettes.xslx)
- M.4 : Fichier CoNLL-U de la prédiction (M.5 Reverdy – Prédiction.conllu) ;
- M.5[80] : Fichier CoNLL-U de la prédiction (M.6 Reverdy – Correction.conllu).

[80] Les étiquettes de partie du discours sont fournies à titre informatif ; elles n'ont pas été corrigées.

A.1 Corpus

POINTE DE L'AILE

Tel que tu es je t'aime tout à fait, disait-il.

Avec ces amateurs d'idéal enchaîné, on arrive vite à la fatigue des appétits éteints. La lune a nourri tant d'hommes qui nous ont légué leur dégoût.

Mais le soleil persiste. Les branches des arbres s'assoupissent quand il fait noir sur la terre.

Quand l'atmosphère se raffermit et vibre, les yeux verts clignotent aux rayons. Le cœur renvoie leur sang aux âmes matérielles. La pierre arrête le sentier rebattu des pas éternels d'une lutte inégale et stérile. Mais devant l'inutile essor le pauvre, souverain de lui-même, gardera la fierté du silence.

ÇA

Les quelques raies qui raccourcissent le mur sont des indications pour la police. Les arbres sont des têtes, ou les têtes des arbres, en tout cas les têtes des arbres me menacent.

Elles courent tout le long du mur et j'ai peur d'arriver à l'endroit où l'on ouvre la grille. Sur la route mon ombre me suit, oblique, et me dit que je cours trop vite. C'est moi qui ai l'air d'un voleur. Enfin, près du petit bois d'où sort le pavillon, je vais crier, je crie, mais des pas tranquilles me rassurent. Et quelqu'un vient m'ouvrir. Par la porte j'aperçois des amis qui sont en train de rire.

Peut-être est-il question de moi ?

D'UNE AUTRE RIVE

Un être qui n'aurait jamais connu son cœur — quelqu'un qui n'en aurait pas l'air.

Il pleure.

— Vous avez brisé mon miroir.

— Pourtant je n'ai fait que crier.

— Vous avez crié trop fort et vous avez brisé mon miroir, les bambous et cette tige encore plus mince que j'aimais. Vous avez brisé son sourire.

La face grimaçante se détourne, et, de l'autre côté de l'eau, une forme très blanche entre les arbres verts qui bougent.

— Elle n’est plus prise dans ton miroir, ni cachée derrière la fumée trop noire de ta pipe.

Relève un peu ta rame et, sur l’eau, allonge les rides mouvantes du sourire.

POURSUITE A TEMPS

La muraille inclinée au son du cor qui bâille — dans le matin tremblant tous les gestes frappés dans les lignes égales de la route aux forêts.

La foule descendant aux foulées droites et, sous les arcs — dans les tendres filets jetés à la traverse du jour bigarré que traversent les élans étouffés du vent, de l’animal.

Sur les revers à peine revenus, aux carrefours émus par le passage, les battements de pieds au sol interrompus, le souffle blanc dans les buissons qui se dispersent.

Tout entière, la région tourne au fil du cadran — les élans des rayons dorés sous la paupière doublés par le bruit sourd de la vie des rivières et des pentes gardées par des plis remuants — jusqu’aux franges du ciel où fument les prières, dans la campagne grise et les cris du couchant.

Le pont jette son dos et l’homme s’agenouille au secours du passant. On revient des échos du monde et de la fête à ce mauvais tournant.

L’ombre est vide sous l’aile qui passe et remue l’air. L’heure sonne pour l’homme que trouble cet espace comblé par son regard.

FAUX PORTRAIT

Dans les bandes limpides du ciel et les rayons fermés, dans les forêts vivantes que le vent soulève et fait tourner, les premières lueurs, les appels des bêtes réveillées.

La campagne s’étire et l’eau se remet à couler.

Dans l’espace ce terrain blanc toujours inoccupé cette tache de sol blafard.

Ce monde de paysan immobile et muet. La figure.

Les yeux et la parole.

Est-ce le calme défendu par des barrières ? Les racines perdues par le temps qui s’oublie ? Une autre direction ou la même aventure ?

Les âmes et les corps à jamais réunis.

LA VOIE DANS LA VILLE

Le grelot de la lune, la pointe du kiosque et la boule du toit.

L’atmosphère tinte.

On annonce la nuit.

Alors on s’aperçoit que les nuages sont enfermés.

Le globe est transparent.

Mais d’en bas on ne voit pas le verre ;

On ne pourrait pas le voir.

Ce soir la pointe du kiosque crève le toit, le verre.

Elle accroche le train qui passe, chargé de têtes et de lampes.

Le boulevard est plein de signes, entre les deux trottoirs ;

et de sourires étouffés près de la bouche. L’été, l’arbre de feu et la tente du cirque.

GLOBE

Où ai-je vu le comédien, le musicien l’homme de Dieu.

Ce n’était qu’un profil qui s’abattait sur la muraille. Une ombre. Nous étions dehors et il pleuvait. Alors mêlées à la pluie on distingua quelques étoiles et un petit enfant tendait sa main.

Quelqu’un cria dans la rue, derrière un volet parce qu’il pleuvait, et tout s’évanouit.

Pas même la nuit, ni l’homme, ni Dieu.

Pas même l’enfant ni les étoiles.

CARACTÈRE SEUL

A pied, courant dans le désert des vents brouillés par le plus sombre caractère, toutes lumières inclinées vers la nuit, les têtes dépouillées des fleurs sentimentales et des sources d’esprit. Et, dans le coin, cette épaisseur de croûte que les jours passent à l’usage des mains.

Cette impalpable agglomération de plans de races, l’os à demi rempli de liquide et remis.

Me reconnaîtrai-je plus tard au milieu de ce carnaval adultère ? Ma main ramenée au départ, la nuque bien tranchée. Je crois. Et je suis si loin du régime animal. La bête relevant ses manches pour prédire. Celle qui tire son orgueil du cri dans l’étang en hiver.

Mélangés dans la même traînée au temps qui se dévide, tirant d’autant.

Avec ceux qui restent perdus dans le désert — la piste désolée, les pierres vides — les vents brouillés — le caractère sombre et inquiétant.

VERS LA FOI

Dans le coin le plus sûr, et sans qu'il y paraisse, la main du songe creux, au coin de l'éventail.

La lumière s'affaisse et sur le mur la tête contre la poitrine et le vitrail.

La tête.

Peut-être la tête de Dieu.

Le vide au feu, partout où la matière manque, la confiance en soi et ceux qui sont autour, mais le rire et l'éclat. Tout sombre et fait le tour.

Cependant quelques visages idiots conservent leur sérieux car, pour eux, rien ne compte, mais eux.

Faut-il avoir assez de confiance pour s'agiter.

Assez de poids léger pour courir sur le vent et suivre ses détours.

Et, toujours d'aplomb, le masque est impassible.

Les girouettes de l'art, de la vie, de la ville.

ET S'EN ALLER

Pendant que les éclairs luisants rayaient l'orage les voix dans les maisons prenaient un autre Son.

La bouche ouverte au vent, la porte que la main pousse et qui se détache, le tourbillon de flamme et l'eau qui tombe à verse — le refrain.

La vitre est éclairée comme un visage — qui se cache et revient. Dans les rideaux, le mouvement du temps et l'esprit qui se lasse quand la pendule saute les heures en écoutant.

Il y a des gens venus de partout et qui parlent — les têtes ramenant l'esprit qui se souvient — et le ciel, qui descend plus lourd sur l'arbre qui se dresse, ouvre une porte basse par où tombe le soir.

Les éclairs sont restés debout sur le fond sombre — les têtes remuées en rond près des rideaux et les visages éclairés contre la vitre — les yeux ouverts qui n'ont jamais fini de regarder.

DERRIÈRE LE CLOCHER

Le jour fume en s'élevant de derrière le toit, l'arbre sec, l'axe mouvant des piles, la colline bordée de cloisons résonnantes.

L’air et l’eau se meuvent par fragments épais qui heurtent leurs éclats et les barrières molles qui les guident.

Au bout du jour, le carrefour tourne vers le couchant ; le coude de l’aiguille aimantée vers le gouffre.

Sous la voûte, les boules qui sortent du clocher roulent dans tous les coins et les heures tracent une lumière au milieu des arcs et des moulures.

Les poissons glissent d’abord sur une seule ligne reflétée dans le ciel, l’oiseau change de sens, la fumée se résigne et la voix du dormeur qui sort du mauvais rêve, au milieu des prairies fumées par le soleil, arrive avec un autre accent, d’une autre sphère, assez haute pour que tout ce qui vit s’arrête, écoute et s’interdise dans le passage des courants où rien n’est clair, ni lourd où tout se mêle, glisse, étincelle à la fois, au même mouvement.

AU POINT DU NORD

Entre les ornières lourdes écrasées de soleil l’animal écumant écarte l’air qui brûle chargé d’odeurs de foins massacrés, de fleurs à peine mortes, encombré du vol aigu des mouches délivrées au bâillement de la ligne des portes.

Les plantes sont encore humides au bord de la chaussée et les racines traînent parmi les pierres et les bêtes mouvantes. On comprendrait que cette eau fût la sueur du ciel ; que les gouttes du front seront les boules au ton de cire et les larmes. Quand le rosier éclate dans la plaie saignante du tournant, à l’orée des précipices infinis du monde déployé. Où finira la charpente de ces bâtiments aériens rouges et bleu de ciel — de ces établissements où la mer vient se rendre et s’apprivoise ? Au climat desséchant de cette pente, à la place de ce poids humide du bas-fond, malgré le mouvement des peupliers inquiets qui se le disent — il ne peut y avoir de place pour le cœur gonflé qui se repent — on tranche la noirceur de l’esprit qui s’envole. Tout est plein de sirop et de quiétude molle. Et trop pur, trop lourd dans cette lumière et le revers qui brille. Je change de moisson. Je regarde d’un œil suppliant le dos des villes.

REMORDS

Je vois le petit apprenti sur l’appareil des rigoles isolées. Je tends la main aux flaques d’eau sous l’éternelle glace perpendiculaire, trouble et où s’évaporent le col, la fissure du treillage chevelu.

Parure de sel, figure de rayons, passage secret des moules de ma main sur les fleurs décapitées, à peine filtrées au réveil, des neiges perdues dans les cimetières, dans les saisons nues, dans le corps ruisselant des larmes du crime muselé. La valse amère.

RÉVEIL INTÉRIEUR

Si le tiroir s'ouvre sous la face béante du meuble c'est le rire ou la bouche du mur.

Les feuilles de soleil se détachent et volent — les carrés d'or se posent et les reflets des glaces se décollent.

Les chiffres qui finissent par prendre aussi de l'épaisseur marquent les coups sur le timbre du ciel qui compte tous les jours.

Le calendrier dehors, le livre ouvert des arbres, les feuilles de soleil se fanent sur le mur ; mais que veulent dire cette main qui cherche, ces yeux qui courent, ces paupières qui battent et ces gouttes d'eau qui restent au bord des lignes du matin ?

AU BORD DES TERRES

A la même minute où les routes repliaient sous la pluie leurs tas de pierres.

Sur la banquette des ressacs et des secousses au tapis vert rayé, au rouge du couchant, le velours effacé et les jambes pendantes ; la fatigue de la mémoire endormie sous la mousse.

Il cherche les états des repas et l'emploi des lumières dans cette nuit mêlée. Rien ne sort. La lune à son filon donne de l'or battu. Les planches du jardin noyées des pleurs de l'arbre. Tous les départs et les voyages commencés interrompus et le ciel dépouillé des plus belles étoiles.

Une bouche édentée bâille — la nuit s'étend — les becs de gaz au fond des abîmes s'alignent — le chemin nouveau des processions marquées sur les trottoirs luisants au bord des retraites marines. Car la mer est partout dans l'ombre où l'on attend — avec son bruit de souffle lourd et ses rires de vagues — toutes ces voix qui nous appellent dans le vent. La peur cachée sort par moments de la nuit noire.

TROIS ÉTOILES

Dans le fond de la fosse au lieu d'un ours sévère — un vieux renard. Ce qui surprend d'abord les touristes venus pour visiter cette guérite de gardien de square au ras du sol et même un peu au-dessous du niveau de la terre — c'est la couleur du ciel, du carré de ciel qui couvre la colline.

Puis ce sont les vêtements masculins de cette étrange femme un peu trop vieille.

Enfin tout est un sujet d'étonnement pour les visiteurs de l'inventeur. Pourtant il faut encore lire la pancarte et laisser des arrhes sur les pourboires dus.

Car tout ici s'entretient de pourboires — même l'animal qui tient les barreaux de la cage — même l'homme qui soigne l'animal.

Male la nuit l'homme, la femme, l'animal — c'est une illusion — il n'y a plus que le marchand de cartons peints qui, sur le seuil, fume sa pipe.

LA PEAU DE NÈGRE

Doucement la poudre de saveur, couleur des fonds tombe sur la moulure égratignée du cadre.

La tempête se déchaîne trop haut pour troubler le calme des buissons, des parcs lumineux et des sombres racines.

La pierre — la pierre bouge en suivant le sabot et c'est l'immensité, la densité, la lourdeur du sol sous le pas.

L'œil se détache peu à peu de cette terre — les dessous humides s'unissent et là où on ne mesure plus la profondeur il y a des rencontres imprévues, des sauts de joie, des regards pleins de haine.

Il nous manque encore du soleil.

Tout le monde est venu dans ce coin, ce recoin. Les plus noirs, les plus mous, les plus vagues sont venus de plus loin — mais… les instruments de la fanfare éclatent, le chef d'orchestre tombe, les fenêtres qui s'épanouissent et les fleurs se noient dans un nouveau silence, car ici il n'y a pas d'autre air.

Le morceau continue.

UNE TACHE SUR LA NUIT

Celui qui descend parmi les avalanches derrière les toits blancs et les arbres qui plient — qui se regarde, s'arrête et tend son bras jusqu'à la paroi de verre qui est peut-être la seule ligne courbe de l'infini.

Celui-là — le corps et la tête et l'âme dans l'espace — tout ce qui dure encore — et traîne sur le soir. Les mots qu'on dit au fond — le bruit confus qui monte — on entend ceux qui parlent à travers les rayons. La fumée à cheval sur l'arête — l'oiseau qui sort la

nuit. Tout est plus grand dehors — les ombres où d'autres formes tombent — les lumières du ciel — la route autour du monde — l'homme seul plus petit.

L'AME ARDENTE

La flamme monte à mesure que le froid s'abaisse sur la nuit.

La flamme de la lampe monte entre les ombres froides qui bougent dans la nuit.

Et la lueur s'allonge et pousse comme un arbre.

Un arbre de feu dans la nuit, sur les routes de glace, entre les parapets de lune et de métal sous les flèches piquantes de mille rayons de cristal ou de reflets d'étoiles.

Vers la flamme qui monte droite dans la nuit.

C'est la voix de la foule obscure qui murmure ou le bruit des pas qui battent le chemin.

Mais jusqu'où poussera la flamme qui monte, ardente et droite, dans la nuit…

L'HOMME AUX ÉTOILES

Une lampe dans chaque main. D'un bout de la chaîne aux étoiles. Les fenêtres bleues du matin, le toit verni et l'escalier qui descend plus bas que la toile. Car il y a la mer entre le mur et l'homme et la nuit dépliée qui arrête le bruit. Il y a le bateau blanc qui écarte les lames et l'aile du soleil qui partage le vent.

Mais, surtout, le front troué par les épines, le cœur d'où sort la flamme et les yeux éplorés — le regard frappe au ciel et la porte qui s'ouvre laisse entrevoir l'espace où remuent les formes mortes sur les chemins tracés par un doigt lumineux.

Les arbres du jardin fermé sont sur la grille — les pointes du signal à côté de la mer — les deux battants ouverts sur l'horizon qui grince — le jour lâché — s'évade et piétine les ombres — les hommes — les étoiles tombées sur le revers.

NUMÉROS VIDES

Je crois me rappeler qu'en comptant bien il n'y avait pas plus de douze numéros sur la façade. Et même je revois le deux un peu plus à droite que les autres. Écrit à la main, une grande main que je n'aurais pas voulu voir reparaître, il semblait vouloir franchir la haie vive du balcon comme un cheval cabré. Le deux ! Tous les autres se détachaient moins bien. Mais il y avait encore le visage immobile derrière le volet et la prunelle bleue qui restait indécise.

Tout à coup une clarté nouvelle détourna l'attention des passants. Une porte s'ouvrait, au fond de l'avenue le ciel se détendait et le vent, qui venait de l'autre côté du sol, faisait flotter les franges des tentures.

Était-ce bien le numéro gagnant. N'y avait-il pas dans cette façon de m'avertir quelque erreur d'écriture. La façade a changé de couleur ; elle tremble au souffle glacé de la nuit et au reflet des lampes qui éclairent trop mal pour calmer cette incertitude.

Douze signes que je ne comprends pas dansent sur le plus large côté du balcon avec les lampes serrées dans les replis mouvants de la voilure.

Le sept, le huit, le neuf. D'où vient que cet ordre m'émeut, moi qui n'ai jamais pu comprendre le sens précis que l'on donnait aux chiffres ?

L'AME EN PÉRIL

Il saute à part, la peur de la mort le regagne.

Ces lignes qui filent dans le creux sans fin et se rejoignent c'est la route de l'imagination sans surprise. La lumière au coin du bois et des manches pleines de vent, la lumière de cette lampe cassée éclaire à peine le front rayé de racines du vieillard studieux qui compte les gouttes de pluie savantes de l'année. La voiture habituelle roule contre les murs qui frissonnent chargée de déserteurs qui passent la frontière du jour. Le cœur s'éprend de cette lumière cassée qui contrarie le vent et la pluie, qui arrête l'infini de l'horreur puissante des tempêtes.

Cette lampe, une goutte de sang dans l'aine de la nuit, détend les muscles et les nerfs du coureur passionné qui a laissé la peur s'emprisonner dans Sa poitrine. Il saute à part, les fossés des raisons sont pleins d'eau. Il fuit le silence hébété, à peine dégagé des rayons lumineux des roues de la tempête — les ornières de sa destinée pleines de vase. Pourtant les courbes de son front gardent l'habitude pressée de l'auréole. Ce n'est pas le bonheur qui manque aux circonvolutions de cette tête. Ce n'est pas la grandeur qui échappe aux idées lugubres de ce nouveau prophète. Mais l'oubli, le dessin mal établi de son séjour d'étranger sans passeport. Une parole aux hommes de son temps aux sources du mépris. Merci. Je ne veux pas de cette gloire.

LES GRAINES DE LA LIBERTÉ

Le bruit du sang se couche lentement sous les feuilles grises des projets qui bordent les rails et la route ; le plein visage du retard sans tenir compte des moindres

encoignures. Un ruisseau d'ébène bruit sous les roches — et le refuge des poissons foudroyants dans le brouillard, l'agonie des lampes et des signes. Le calme des rides sanglantes et violacées par le froid matinal, les brûlures du vent déchiré aux épines des cataclysmes, quand les troupeaux de cœurs gonflés rentrent trop tard. Quand les heures passées ensemble se regrettent. Quand se quittent les silences vécus en tête à tête. Les jeux de cartes, les disputes de l'amour-propre au domino, je suis las de vivre dans la nuit des jours meilleurs. Je cherche les journaux dans la toile des lignes, les bustes des photographies et des paysages, le monde gémissant écrasé sous le poids des soupirs, des efforts et des malheurs construits à la truelle. Alors la suie des papillons, qui ont fini de tromper la lumière, s'accorde aux doigts des feuilles des mourants. Les cristaux s'écartent sur la nuit. Les poumons de la liberté boivent de l'encre.

ÉTOILE FILANTE

Il y a des éclairs sans cigales à l'horizon, il y a des déchirures sans une goutte de sang à la tempête, mais il y a surtout, dans le désert et la raison sans oasis, les fenêtres sans abri, les lumières sans écran, les abat-jour sans rayons obliques de la guillotine qui exécute les plus criminels souvenirs de mon temps de forçat.

Où avez-vous dérobé cette photographie sans rayures de ma misérable personnalité morale et de mon corps aux contours si bourgeois. Par toutes les épines qui sillonnent les routes du cœur et de la pensée, chemins coupés de meurtrissures, de rives d'eau, de colliers de larmes et de signes, tracés par la haine et le ressentiment des bêtes, je ne me reconnais pas dans ces pages au miroir méfiant de la source.

Dans cette flaque d'eau où tremble le péril, où le temps s'accumule au goutte à goutte, défiant les menaces du ciel.

Je suis un témoignage fendu de la tête aux pieds, une indication précise, mais fugitive de ce qu'a voulu dire la création en remontant de nos jours jusqu'au commencement des termes.

LA TÊTE PLEINE DE BEAUTÉ

Dans l'abîme doré, rouge, glacé, doré, l'abîme où gîte la douleur, les tourbillons roulants entraînent les bouillons de mon sang dans les vases, dans les retours de flammes de mon tronc. La tristesse moirée s'engloutit dans les crevasses tendres du cœur. Il y a des accidents obscurs et compliqués, impossibles à dire. Et il y a pourtant

l’esprit de l’ordre, l’esprit régulier, l’esprit commun à tous les désespoirs qui interroge. O toi qui traines sur la vie, entre les buissons fleuris et pleins d’épines de la vie, parmi les feuilles mortes, les reliefs de triomphes, les appels sans secours, les balayures mordorées, la poudre sèche des espoirs, les braises noircies de la gloire, et les coups de révolte, toi, qui ne voudrais plus désormais aboutir nulle part. Toi, source intarissable de sang. Toi, désastre intense de lueurs qu’aucun jet de source, qu’aucun glacier rafraîchissant ne tentera jamais d’éteindre de sa sève. Toi, lumière. Toi, sinuosité de l’amour enseveli qui se dérobe. Toi, parure des ciels cloués sur les poutres de l’infini. Plafond des idées contradictoires. Vertigineuse pesée des forces ennemies. Chemins mêlés dans le fracas des chevelures. Toi, douceur et haine — horizon ébréché, ligne pure de l’indifférence et de l’oubli. Toi, ce matin, tout seul dans l’ordre, le calme, et la révolution universelle. Toi, clou de diamant. Toi, pureté, pivot éblouissant du flux et du reflux de ma pensée dans les lignes du monde.

A.2 – Abréviations des titres des poèmes

Pointe de l'aile	PA	**Réveil intérieur**	RI
Ça	C	**Au bord des terres**	BT
D'une autre rive	AR	**Trois étoiles**	TE
Poursuite à temps	PT	**La peau de nègre**	LPN
Faux portrait	FP	**Une tache sur la nuit**	TN
La voie dans la ville	VV	**L'âme ardente**	AA
Globe	G	**L'homme aux étoiles**	HE
Caractère seule	CS	**Numéros vides**	NV
Vers la foi	VF	**L'âme en péril**	AP
Et s'en aller	SA	**Les graines de la liberté**	GL
Derrière le clocher	DC	**Étoile filante**	EF
Au point du nord	APN	**La tête pleine de beauté**	TPB
Remords	R		

A.3 – Phrases de notre corpus

(PA-1) Tel que tu es je t'aime tout à fait, disait-il.

(PA-2) Avec ces amateurs d'idéal enchaîné, on arrive vite à la fatigue des appétits éteints.

(PA-3) La lune a nourri tant d'hommes qui nous ont légué leur dégoût.

(PA-4), Mais le soleil persiste.

(PA-5) Les branches des arbres s'assoupissent quand il fait noir sur la terre.

(PA-6) Quand l'atmosphère se raffermit et vibre, les yeux verts clignotent aux rayons.

(PA-7) Le cœur renvoie leur sang aux âmes matérielles.

(PA-8) La pierre arrête le sentier rebattu des pas éternels d'une lutte inégale et stérile.

(PA-9), Mais devant l'inutile essor le pauvre, souverain de lui-même, gardera la fierté du silence.

(C-1) Les quelques raies qui raccourcissent le mur sont des indications pour la police.

(C-2) Les arbres sont des têtes, ou les têtes des arbres, en tout cas les têtes des arbres me menacent.

(C-3) Elles courent tout le long du mur et j'ai peur d'arriver à l'endroit où l'on ouvre la grille.

(C-4) Sur la route mon ombre me suit, oblique, et me dit que je cours trop vite.

(C-5) C'est moi qui ai l'air d'un voleur.

(C-6) Enfin, près du petit bois d'où sort le pavillon, je vais crier, je crie, mais des pas tranquilles me rassurent.

(C-7) Et quelqu'un vient m'ouvrir.

(C-8) Par la porte j'aperçois des amis qui sont en train de rire.

(C-9) Peut-être est-il question de moi ?

(AR-1) Un être qui n'aurait jamais connu son cœur — quelqu'un qui n'en aurait pas l'air.

(AR-2) Il pleure.

(AR-3) — Vous avez brisé mon miroir.

(AR-4) — Pourtant je n'ai fait que crier.

(AR-5) — Vous avez crié trop fort et vous avez brisé mon miroir, les bambous et cette tige encore plus mince que j'aimais.

(AR-6) Vous avez brisé son sourire.

(AR-7) La face grimaçante se détourne, et, de l'autre côté de l'eau, une forme très blanche entre les arbres verts qui bougent.

(AR-8) — Elle n'est plus prise dans ton miroir, ni cachée derrière la fumée trop noire de ta pipe.

(AR-9) Relève un peu ta rame et, sur l'eau, allonge les rides mouvantes du sourire.

(PT-1) La muraille inclinée au son du cor qui bâille — dans le matin tremblant tous les gestes frappés dans les lignes égales de la route aux forêts.

(PT-2) La foule descendant aux foulées droites et, sous les arcs — dans les tendres filets jetés à la traverse du jour bigarré que traversent les élans étouffés du vent, de l'animal.

(PT-3) Sur les revers à peine revenus, aux carrefours émus par le passage, les battements de pieds au sol interrompus, le souffle blanc dans les buissons qui se dispersent.

(PT-4) Tout entière, la région tourne au fil du cadran — les élans des rayons dorés sous la paupière doublés par le bruit sourd de la vie des rivières et des pentes gardées par des plis remuants — jusqu'aux franges du ciel où fument les prières, dans la campagne grise et les cris du couchant.

(PT-5) Le pont jette son dos et l'homme s'agenouille au secours du passant.

(PT-6) On revient des échos du monde et de la fête à ce mauvais tournant.

(PT-7) L'ombre est vide sous l'aile qui passe et remue l'air.

(PT-8) L'heure sonne pour l'homme que trouble cet espace comblé par son regard.

(FP-1) Dans les bandes limpides du ciel et les rayons fermés, dans les forêts vivantes que le vent soulève et fait tourner, les premières lueurs, les appels des bêtes réveillées.

(FP-2) La campagne s'étire et l'eau se remet à couler.

(FP-3) Dans l'espace ce terrain blanc toujours inoccupé cette tache de sol blafard.

(FP-4) Ce monde de paysan immobile et muet.

(FP-5) La figure.

(FP-6) Les yeux et la parole.

(FP-7) Est-ce le calme défendu par des barrières ?

(FP-8) Les racines perdues par le temps qui s'oublie ?

(FP-9) Une autre direction ou la même aventure ?

(FP-10) Les âmes et les corps à jamais réunis.

(VV-1) Le grelot de la lune, la pointe du kiosque et la boule du toit.

(VV-2) L'atmosphère tinte.

(VV-3) On annonce la nuit.

(VV-4) Alors on s'aperçoit que les nuages sont enfermés.

(VV-5) Le globe est transparent.

(VV-6), Mais d'en bas on ne voit pas le verre ; on ne pourrait pas le voir.

(VV-7) Ce soir la pointe du kiosque crève le toit, le verre.

(VV-8) Elle accroche le train qui passe, chargé de têtes et de lampes.

(VV-9) Le boulevard est plein de signes, entre les deux trottoirs ; et de sourires étouffés près de la bouche.

(VV-10) L'été, l'arbre de feu et la tente du cirque.

(G-1) Où ai-je vu le comédien, le musicien l'homme de Dieu.

(G-2) Ce n'était qu'un profil qui s'abattait sur la muraille.

(G-3) Une ombre.

(G-4) Nous étions dehors et il pleuvait.

(G-5) Alors mêlées à la pluie on distingua quelques étoiles et un petit enfant tendait sa main.

(G-6) Quelqu'un cria dans la rue, derrière un volet parce qu'il pleuvait, et tout s'évanouit.

(G-7) Pas même la nuit, ni l'homme, ni Dieu.

(G-8) Pas même l'enfant ni les étoiles.

(CS-1) À pied, courant dans le désert des vents brouillés par le plus sombre caractère, toutes lumières inclinées vers la nuit, les têtes dépouillées des fleurs sentimentales et des sources d'esprit.

(CS-2) Et, dans le coin, cette épaisseur de croûte que les jours passent à l'usage des mains.

(CS-3) Cette impalpable agglomération de plans de races, l'os à demi rempli de liquide et remis.

(CS-4) Me reconnaîtrai-je plus tard au milieu de ce carnaval adultère ?

(CS-5) Ma main ramenée au départ, la nuque bien tranchée.

(CS-6) Je crois.

(CS-7) Et je suis si loin du régime animal.

(CS-8) La bête relevant ses manches pour prédire.

(CS-9) Celle qui tire son orgueil du cri dans l'étang en hiver.

(CS-10) Mélangés dans la même traînée au temps qui se dévide, tirant d'autant.

(CS-11) Avec ceux qui restent perdus dans le désert — la piste désolée, les pierres vides — les vents brouillés — le caractère sombre et inquiétant.

(VF-1) Dans le coin le plus sûr, et sans qu'il y paraisse, la main du songe creux, au coin de l'éventail.

(VF-2) La lumière s'affaisse et sur le mur la tête contre la poitrine et le vitrail.

(VF-3) La tête.

(VF-4) Peut-être la tête de Dieu.

(VF-5) Le vide au feu, partout où la matière manque, la confiance en soi et ceux qui sont autour, mais le rire et l'éclat.

(VF-6) Tout sombre et fait le tour.

(VF-7) Cependant quelques visages idiots conservent leur sérieux car, pour eux, rien ne compte, mais eux.

(VF-8) Faut-il avoir assez de confiance pour s'agiter.

(VF-9) Assez de poids léger pour courir sur le vent et suivre ses détours.

(VF-10) Et, toujours d'aplomb, le masque est impassible.

(VF-11) Les girouettes de l'art, de la vie, de la ville.

(SA-1) Pendant que les éclairs luisants rayaient l'orage les voix dans les maisons prenaient un autre son.

(SA-2) La bouche ouverte au vent, la porte que la main pousse et qui se détache, le tourbillon de flamme et l'eau qui tombe à verse — le refrain.

(SA-3) La vitre est éclairée comme un visage — qui se cache et revient.

(SA-4) Dans les rideaux, le mouvement du temps et l'esprit qui se lasse quand la pendule saute les heures en écoutant.

(SA-5) Il y a des gens venus de partout et qui parlent — les têtes ramenant l'esprit qui se souvient — et le ciel, qui descend plus lourd sur l'arbre qui se dresse, ouvre une porte basse par où tombe le soir.

(SA-6) Les éclairs sont restés debout sur le fond sombre — les têtes remuées en rond près des rideaux et les visages éclairés contre la vitre — les yeux ouverts qui n'ont jamais fini de regarder.

(DC-1) Le jour fume en s'élevant de derrière le toit, l'arbre sec, l'axe mouvant des piles, la colline bordée de cloisons résonnantes.

(DC-2) L'air et l'eau se meuvent par fragments épais qui heurtent leurs éclats et les barrières molles qui les guident.

(DC-3) Au bout du jour, le carrefour tourne vers le couchant ; le coude de l'aiguille aimantée vers le gouffre.

(DC-4) Sous la voûte, les boules qui sortent du clocher roulent dans tous les coins et les heures tracent une lumière au milieu des arcs et des moulures.

(DC-5) Les poissons glissent d'abord sur une seule ligne reflétée dans le ciel, l'oiseau change de sens, la fumée se résigne et la voix du dormeur qui sort du mauvais rêve, au milieu des prairies fumées par le soleil, arrive avec un autre accent, d'une autre sphère, assez haute pour que tout ce qui vit s'arrête, écoute et s'interdise dans le passage des courants où rien n'est clair, ni lourd où tout se mêle, glisse, étincelle à la fois, au même mouvement.

(APN-1) Entre les ornières lourdes écrasées de soleil l'animal écumant écarte l'air qui brûle chargé d'odeurs de foins massacrés, de fleurs à peine mortes, encombré du vol aigu des mouches délivrées au bâillement de la ligne des portes.

(APN-2) Les plantes sont encore humides au bord de la chaussée et les racines traînent parmi les pierres et les bêtes mouvantes.

(APN-3) On comprendrait que cette eau fût la sueur du ciel ; que les gouttes du front seront les boules au ton de cire et les larmes.

(APN-4) Quand le rosier éclate dans la plaie saignante du tournant, à l'orée des précipices infinis du monde déployé.

(APN-5) Où finira la charpente de ces bâtiments aériens rouges et bleu de ciel — de ces établissements où la mer vient se rendre et s'apprivoise ?

(APN-6) Au climat desséchant de cette pente, à la place de ce poids humide du bas-fond, malgré le mouvement des peupliers inquiets qui se le disent — il ne peut y avoir de place pour le cœur gonflé qui se repent — on tranche la noirceur de l'esprit qui s'envole.

(APN-7) Tout est plein de sirop et de quiétude molle.

(APN-8) Et trop pur, trop lourd dans cette lumière et le revers qui brille.

(APN-9) Je change de moisson.

(APN-10) Je regarde d'un œil suppliant le dos des villes.

(R-1) Je vois le petit apprenti sur l'appareil des rigoles isolées.

(R-2) Je tends la main aux flaques d'eau sous l'éternelle glace perpendiculaire, trouble et où s'évaporent le col, la fissure du treillage chevelu.

(R-3) Parure de sel, figure de rayons, passage secret des moules de ma main sur les fleurs décapitées, à peine filtrées au réveil, des neiges perdues dans les cimetières, dans les saisons nues, dans le corps ruisselant des larmes du crime muselé.

(R-4) La valse amère.

(RI-1) Si le tiroir s'ouvre sous la face béante du meuble c'est le rire ou la bouche du mur.

(RI-2) Les feuilles de soleil se détachent et volent — les carrés d'or se posent et les reflets des glaces se décollent.

(RI-3) Les chiffres qui finissent par prendre aussi de l'épaisseur marquent les coups sur le timbre du ciel qui compte tous les jours.

(RI-4) Le calendrier dehors, le livre ouvert des arbres, les feuilles de soleil se fanent sur le mur ; mais que veulent dire cette main qui cherche, ces yeux qui courent, ces paupières qui battent et ces gouttes d'eau qui restent au bord des lignes du matin ?

(BT-1) A la même minute où les routes repliaient sous la pluie leurs tas de pierres.

(BT-2) Sur la banquette des ressacs et des secousses au tapis vert rayé, au rouge du couchant, le velours effacé et les jambes pendantes ; la fatigue de la mémoire endormie sous la mousse.

(BT-3) Il cherche les états des repas et l'emploi des lumières dans cette nuit mêlée.

(BT-4) Rien ne sort.

(BT-5) La lune à son filon donne de l'or battu.

(BT-6) Les planches du jardin noyées des pleurs de l'arbre.

(BT-7) Tous les départs et les voyages commencés interrompus et le ciel dépouillé des plus belles étoiles.

(BT-8) Une bouche édentée bâille — la nuit s'étend — les becs de gaz au fond des abîmes s'alignent — le chemin nouveau des processions marquées sur les trottoirs luisants au bord des retraites marines.

(BT-9), Car la mer est partout dans l'ombre où l'on attend — avec son bruit de souffle lourd et ses rires de vagues — toutes ces voix qui nous appellent dans le vent.

(BT-10) La peur cachée sort par moments de la nuit noire.

(BT-11) Dans le fond de la fosse au lieu d'un ours sévère — un vieux renard.

(TE-1) Ce qui surprend d'abord les touristes venus pour visiter cette guérite de gardien de square au ras du sol et même un peu au-dessous du niveau de la terre — c'est la couleur du ciel, du carré de ciel qui couvre la colline.

(TE-2) Puis ce sont les vêtements masculins de cette étrange femme un peu trop vieille.

(TE-3) Enfin tout est un sujet d'étonnement pour les visiteurs de l'inventeur.

(TE-4) Pourtant il faut encore lire la pancarte et laisser des arrhes sur les pourboires dus.

(TE-5), Car tout ici s'entretient de pourboires — même l'animal qui tient les barreaux de la cage — même l'homme qui soigne l'animal.

(TE-6) Male la nuit l'homme, la femme, l'animal — c'est une illusion — il n'y a plus que le marchand de cartons peints qui, sur le seuil, fume sa pipe.

(LPN-1) Doucement la poudre de saveur, couleur des fonds tombe sur la moulure égratignée du cadre.

(LPN-2) La tempête se déchaîne trop haut pour troubler le calme des buissons, des parcs lumineux et des sombres racines.

(LPN-3) La pierre — la pierre bouge en suivant le sabot et c'est l'immensité, la densité, la lourdeur du sol sous le pas.

(LPN-4) L'œil se détache peu à peu de cette terre — les dessous humides s'unissent et là où on ne mesure plus la profondeur il y a des rencontres imprévues, des sauts de joie, des regards pleins de haine.

(LPN-5) Il nous manque encore du soleil.

(LPN-6) Tout le monde est venu dans ce coin, ce recoin.

(LPN-7) Les plus noirs, les plus mous, les plus vagues sont venus de plus loin — mais… les instruments de la fanfare éclatent, le chef d'orchestre tombe, les fenêtres qui s'épanouissent et les fleurs se noient dans un nouveau silence, Car ici il n'y a pas d'autre air.

(LPN-8) Le morceau continue.

(TN-1) Celui qui descend parmi les avalanches derrière les toits blancs et les arbres qui plient — qui se regarde, s'arrête et tend son bras jusqu'à la paroi de verre qui est peut-être la seule ligne courbe de l'infini.

(TN-2) Celui-là — le corps et la tête et l'âme dans l'espace — tout ce qui dure encore — et traîne sur le soir.

(TN-3) Les mots qu'on dit au fond — le bruit confus qui monte — on entend ceux qui parlent à travers les rayons.

(TN-4) La fumée à cheval sur l'arête — l'oiseau qui sort la nuit.

(TN-5) Tout est plus grand dehors — les ombres où d'autres formes tombent — les lumières du ciel — la route autour du monde — l'homme seul plus petit.

(AA-1) La flamme monte à mesure que le froid s'abaisse sur la nuit.

(AA-2) La flamme de la lampe monte entre les ombres froides qui bougent dans la nuit.

(AA-3) Et la lueur s'allonge et pousse comme un arbre.

(AA-4) Un arbre de feu dans la nuit, sur les routes de glace, entre les parapets de lune et de métal sous les flèches piquantes de mille rayons de cristal ou de reflets d'étoiles.

(AA-5) Vers la flamme qui monte droite dans la nuit…

(AA-6) C'est la voix de la foule obscure qui murmure ou le bruit des pas qui battent le chemin.

(AA-7), Mais jusqu'où poussera la flamme qui monte, ardente et droite, dans la nuit.

(HE-1) Une lampe dans chaque main.

(HE-2) D'un bout de la chaîne aux étoiles.

(HE-3) Les fenêtres bleues du matin, le toit verni et l'escalier qui descend plus bas que la toile.

(HE-4), Car il y a la mer entre le mur et l'homme et la nuit dépliée qui arrête le bruit.

(HE-5) Il y a le bateau blanc qui écarte les lames et l'aile du soleil qui partage le vent.

(HE-6), Mais, surtout, le front troué par les épines, le cœur d'où sort la flamme et les yeux éplorés — le regard frappe au ciel et la porte qui s'ouvre laisse entrevoir l'espace où remuent les formes mortes sur les chemins tracés par un doigt lumineux.

(HE-7) Les arbres du jardin fermé sont sur la grille — les pointes du signal à côté de la mer — les deux battants ouverts sur l'horizon qui grince — le jour lâché — s'évade et piétine les ombres — les hommes — les étoiles tombées sur le revers.

(NV-1) Je crois me rappeler qu'en comptant bien il n'y avait pas plus de douze numéros sur la façade.

(NV-2) Et même je revois le deux un peu plus à droite que les autres.

(NV-3) Écrit à la main, une grande main que je n'aurais pas voulu voir reparaître, il semblait vouloir franchir la haie vive du balcon comme un cheval cabré.

(NV-4) Le deux !

(NV-5) Tous les autres se détachaient moins bien.

(NV-6), Mais il y avait encore le visage immobile derrière le volet et la prunelle bleue qui restait indécise.

(NV-7) Tout à coup une clarté nouvelle détourna l'attention des passants.

(NV-8) Une porte s'ouvrait, au fond de l'avenue le ciel se détendait et le vent, qui venait de l'autre côté du sol, faisait flotter les franges des tentures.

(NV-9) Était-ce bien le numéro gagnant.

(NV-10) N'y avait-il pas dans cette façon de m'avertir quelque erreur d'écriture.

(NV-11) La façade a changé de couleur ; elle tremble au souffle glacé de la nuit et au reflet des lampes qui éclairent trop mal pour calmer cette incertitude.

(NV-12) Douze signes que je ne comprends pas dansent sur le plus large côté du balcon avec les lampes serrées dans les replis mouvants de la voilure.

(NV-13) Le sept, le huit, le neuf.

(NV-14) D'où vient que cet ordre m'émeut, moi qui n'ai jamais pu comprendre le sens précis que l'on donnait aux chiffres ?

(AP-1) Il saute à part, la peur de la mort le regagne.

(AP-2) Ces lignes qui filent dans le creux sans fin et se rejoignent c'est la route de l'imagination sans surprise.

(AP-3) La lumière au coin du bois et des manches pleines de vent, la lumière de cette lampe cassée éclaire à peine le front rayé de racines du vieillard studieux qui compte les gouttes de pluie savantes de l'année.

(AP-4) La voiture habituelle roule contre les murs qui frissonnent chargée de déserteurs qui passent la frontière du jour.

(AP-5) Le cœur s'éprend de cette lumière cassée qui contrarie le vent et la pluie, qui arrête l'infini de l'horreur puissante des tempêtes.

(AP-6) Cette lampe, une goutte de sang dans l'aine de la nuit, détend les muscles et les nerfs du coureur passionné qui a laissé la peur s'emprisonner dans Sa poitrine.

(AP-7) Il saute à part, les fossés des raisons sont pleins d'eau.

(AP-8) Il fuit le silence hébété, à peine dégagé des rayons lumineux des roues de la tempête — les ornières de sa destinée pleines de vase.

(AP-9) Pourtant les courbes de son front gardent l'habitude pressée de l'auréole.

(AP-10) Ce n'est pas le bonheur qui manque aux circonvolutions de cette tête.

(AP-11) Ce n'est pas la grandeur qui échappe aux idées lugubres de ce nouveau prophète.

(AP-12), Mais l'oubli, le dessin mal établi de son séjour d'étranger sans passeport.

(AP-13) Une parole aux hommes de son temps aux sources du mépris.

(AP-14) Merci.

(AP-15) Je ne veux pas de cette gloire.

(GL-1) Le bruit du sang se couche lentement sous les feuilles grises des projets qui bordent les rails et la route ; le plein visage du retard sans tenir compte des moindres encoignures.

(GL-2) Un ruisseau d'ébène bruit sous les roches — et le refuge des poissons foudroyants dans le brouillard, l'agonie des lampes et des signes.

(GL-3) Le calme des rides sanglantes et violacées par le froid matinal, les brûlures du vent déchiré aux épines des cataclysmes, quand les troupeaux de cœurs gonflés rentrent trop tard.

(GL-4) Quand les heures passées ensemble se regrettent.

(GL-5) Quand se quittent les silences vécus en tête à tête.

(GL-6) Les jeux de cartes, les disputes de l'amour-propre au domino, je suis las de vivre dans la nuit des jours meilleurs.

(GL-7) Je cherche les journaux dans la toile des lignes, les bustes des photographies et des paysages, le monde gémissant écrasé sous le poids des soupirs, des efforts et des malheurs construits à la truelle.

(GL-8) Alors la suie des papillons, qui ont fini de tromper la lumière, s'accorde aux doigts des feuilles des mourants.

(GL-9) Les cristaux s'écartent sur la nuit.

(GL-10) Les poumons de la liberté boivent de l'encre.

(EF-1) Il y a des éclairs sans cigales à l'horizon, il y a des déchirures sans une goutte de sang à la tempête, mais il y a surtout, dans le désert et la raison sans oasis, les fenêtres sans abri, les lumières sans écran, les abat-jour sans rayons obliques de la guillotine qui exécute les plus criminels souvenirs de mon temps de forçat.

(EF-2) Où avez-vous dérobé cette photographie sans rayures de ma misérable personnalité morale et de mon corps aux contours si bourgeois.

(EF-3) Par toutes les épines qui sillonnent les routes du cœur et de la pensée, chemins coupés de meurtrissures, de rives d'eau, de colliers de larmes et de signes, tracés par la haine et le ressentiment des bêtes, je ne me reconnais pas dans ces pages au miroir méfiant de la source.

(EF-4) Dans cette flaque d'eau où tremble le péril, où le temps s'accumule au goutte à goutte, défiant les menaces du ciel.

(EF-5) Je suis un témoignage fendu de la tête aux pieds, une indication précise, mais fugitive de ce qu'a voulu dire la création en remontant de nos jours jusqu'au commencement des termes.

(TPB-1) Dans l'abîme doré, rouge, glacé, doré, l'abîme où gîte la douleur, les tourbillons roulants entraînent les bouillons de mon sang dans les vases, dans les retours de flammes de mon tronc.

(TPB-2) La tristesse moirée s'engloutit dans les crevasses tendres du cœur.

(TPB-3) Il y a des accidents obscurs et compliqués, impossibles à dire.

(TPB-4) Et il y a pourtant l'esprit de l'ordre, l'esprit régulier, l'esprit commun à tous les désespoirs qui interroge.

(TPB-5) O toi qui traines sur la vie, entre les buissons fleuris et pleins d'épines de la vie, parmi les feuilles mortes, les reliefs de triomphes, les appels sans secours, les balayures mordorées, la poudre sèche des espoirs, les braises noircies de la

gloire, et les coups de révolte, toi, qui ne voudrais plus désormais aboutir nulle part.

(TPB-6) Toi, source intarissable de sang.

(TPB-7) Toi, désastre intense de lueurs qu'aucun jet de source, qu'aucun glacier rafraîchissant ne tentera jamais d'éteindre de sa sève.

(TPB-8) Toi, lumière.

(TPB-9) Toi, sinuosité de l'amour enseveli qui se dérobe.

(TPB-10) Toi, parure des ciels cloués sur les poutres de l'infini.

(TPB-11) Plafond des idées contradictoires.

(TPB-12) Vertigineuse pesée des forces ennemies.

(TPB-13) Chemins mêlés dans le fracas des chevelures.

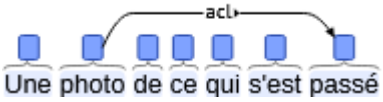
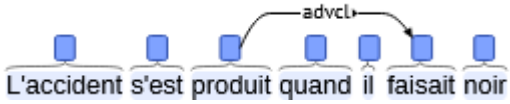
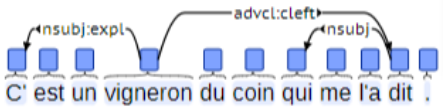
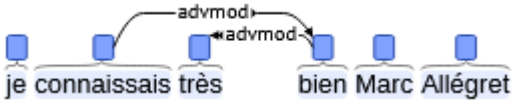
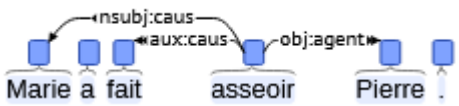
(TPB-14) Toi, douceur et haine — horizon ébréché, ligne pure de l'indifférence et de l'oubli.

(TPB-15) Toi, ce matin, tout seul dans l'ordre, le calme, et la révolution universelle.

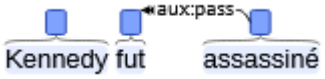
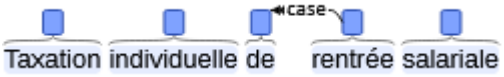
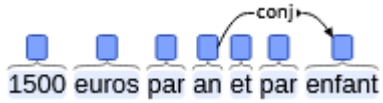
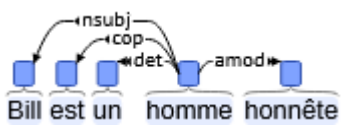
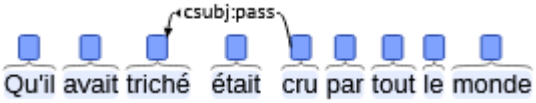
(TPB-16) Toi, clou de diamant.

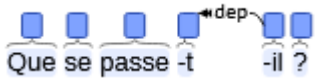
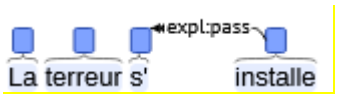
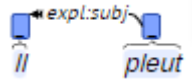
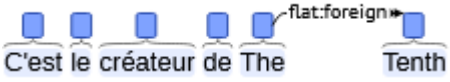
(TPB-17) Toi, pureté, pivot éblouissant du flux et du reflux de ma pensée dans les lignes du monde.

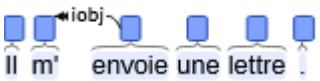
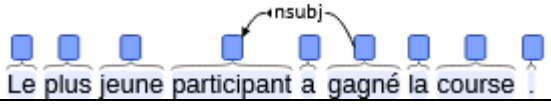
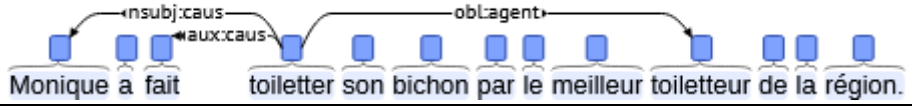
A.4 – Détail et illustration des étiquettes utilisées dans le treebank UD-Sequoia[81]

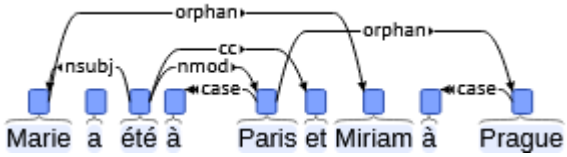
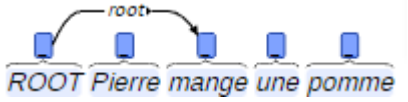
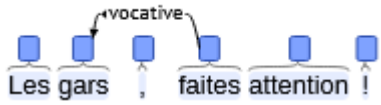
acl *adnominal clause*	Proposition à un mode fini ou non fini qui modifie un substantif. 
acl:relcl *relative clause modifier*	Désigne les propositions relatives. 
advcl, advcl:cleft *adverbial clause modifier*	Proposition adverbiale modifiant un verbe ou un autre prédicat sans être argument actanciel.  *advcl:cleft* est utilisé pour les phrases clivées. 
advmod *adverbial modifier*	Adverbe qui modifie un prédicat ou un autre modificateur. 
amod *adjectival modifier*	Groupe adjectival qui modifie un nom ou un pronom. 
appos *appositional modifier*	Groupe nominal suivant immédiatement un premier nom et qui définit ou modifie ce premier. 
aux:caus *causative auxiliary*	Lie le verbe à l'infinitif et l'auxiliaire causatif (faire) dans une construction causative. 

[81] Les définitions et exemples sont adaptés ou repris de https://universaldependencies.org/u/dep/index.html.

aux:pass *passive auxiliary*	Verbe auxiliaire de la voix passive. 
aux:tense *tense auxiliary*	Verbe auxiliaire servant à la construction des temps. 
case *case marking*	Élément marquant le cas traité comme un mot syntaxique, incluant les prépositions, postpositions et clitiques. 
cc *coordinating conjonction*	Relation entre un élément introduit par une conjonction de coordination et cette conjonction. 
ccomp *clausal complement*	Proposition complément d'un verbe ou d'un adjectif à statut actanciel. 
conj *conjuct*	Relation entre deux éléments connectés par une conjonction de coordination. 
cop *copula*	Marque le mot utilisé (généralement un verbe, mais cela peut être un pronom dans certaines langues) pour lier un sujet à un prédicat non verbal (notamment dans les constructions à verbe d'état). Les verbes copules sont étiquetés *aux* plutôt que *verb*. 
csubj *clausal subject*	Proposition ayant la fonction de sujet syntaxique d'une autre proposition. 
csubj:pass *clausal passive subject*	Proposition ayant la fonction de sujet syntaxique d'une proposition à la voix passive. 

dep *unspecified dependency*	Relation qu'il est impossible de mieux préciser avec les étiquettes disponibles. 
det *determiner*	Relation entre une tête nominale et son déterminant. 
discourse *discourse element*	Relation utilisée pour les interjections et particules du discours qui ne sont pas clairement liées à la structure de la phrase sauf selon un mode expressif ou pragmatique. 
dislocated *dislocated elements*	Relation utilisée pour les éléments préposés ou postposés inscrits dans une structure de reprise syntaxique ; construction courante à l'oral. 
expl:comp *reflexive pronoun used as complement*	Pronom réflexif occupant la fonction de complément actanciel. 
expl:pass *reflexive pronoun used in reflexive passive*	Relation utilisée dans les constructions réflexives passives. 
expl:subj *reflexive pronoun used as subject*	Pronom occupant la place de sujet dans une construction impersonnelle ou une construction interrogative répétant le sujet. 
fixed *fixed*	Relation utilisée dans certaines expressions grammaticalisées à mots multiples qui fonctionnent comme des mots fonctionnels ou des adverbes. 
flat:foreign *foreign words*	Relation qui permet d'étiqueter une séquence de mots en langue étrangère. 
flat:name *names*	Relation qui permet d'étiqueter une expression à mots multiples correspondant à un nom propre.

	
goeswith *goes with*	Relation entre deux ou plusieurs parts séparées dans un texte alors qu'elles devraient être écrites en un seul mot. 
iobj *indirect object*	Complément actanciel correspondant au complément d'objet indirect lorsqu'il est pronominal. Le complément d'objet indirect est réalisé avec une proposition, il est analysé avec la relation *obl:arg*. 
mark *mark*	Marque le mot indiquant qu'une proposition est subordonnée d'une autre, la conjonction de subordination en français. 
nmod *nominal modifier*	Dépendant nominal d'un autre substantif ou groupe nominal qui correspond à la fonction d'attribut ou de complément déterminatif. 
nsubj *Nominal subject*	Substantif ou groupe nominal occupant la fonction de sujet syntaxique d'une proposition. 
nsubj:caus *causative nominal subject*	Sujet syntaxique d'une construction causative (voir aux:caus).
nsubj:pass *passive nominal subject*	Sujet syntaxique d'une proposition à la voix passive. 
nummod *numeric modifier*	Nombre modifiant le sens d'un nom par une quantité. 
obl:agent *agent modifier*	Agent des verbes passifs ou sujet de l'infinitif dans les constructions causatives. 

obl:mod **obl:arg** *oblique modifier or argument*	Relation indiquant un dépendant nominal d'un verbe qui n'est ni sujet (nsubj) ni objet direct (obj) ; obl:mod est utilisé lorsque le dépendant est modifieur du verbe et obl:arg lorsqu'il est argument actanciel. 
obj *object*	Relation avec l'argument actanciel correspondant au COD. 
orphan *remnant in ellipsis*	Relation utilisée pour analyser les cas d'ellipse dans lesquels aucun mot fonctionnel ne peut être utilisé pour prendre la place du mot lexical élidé. 
parataxis *parataxis*	Relation entre un mot et d'autres éléments placés côte à côte sans coordination, subordination ou relation explicite avec la tête. 
punct *punctuation*	Relation à toute marque de ponctuation dans une phrase lorsqu'elle est considérée. 
root *root*	Relation pointant vers la racine de la phrase ; un nœud factice ROOT d'index 0 est utilisé comme gouverneur. 
vocative *vocative*	Adresse à un interlocuteur dans un texte. 
xcomp *open clausal complement*	Complément prédicatif d'un verbe ou un adjectif sans sujet propre. La référence au sujet est nécessairement déterminée par un argument externe au xcomp. 

A.5 – Statistiques de similarité entre Sequoia-Train, Sequoia-Test et Reverdy

Longueur moyenne des phrases

	Sequoia-Train	Sequoia-Test	Reverdy original (segmenté)	Reverdy corrigé (joint)
Moyenne du nombre de tokens par phrase	23,25	22,71	17,69	19,88

Moyenne du nombre de tokens par phrase dans les corpus

Moyenne de la longueur des dépendances (MLD)

	Sequoia-Train	Sequoia-Test	Reverdy original	Reverdy corrigé segmenté	Reverdy corrigé joint
Moyenne de la longueur des dépendances	4,29	4,05	3,77	3,73	3,90

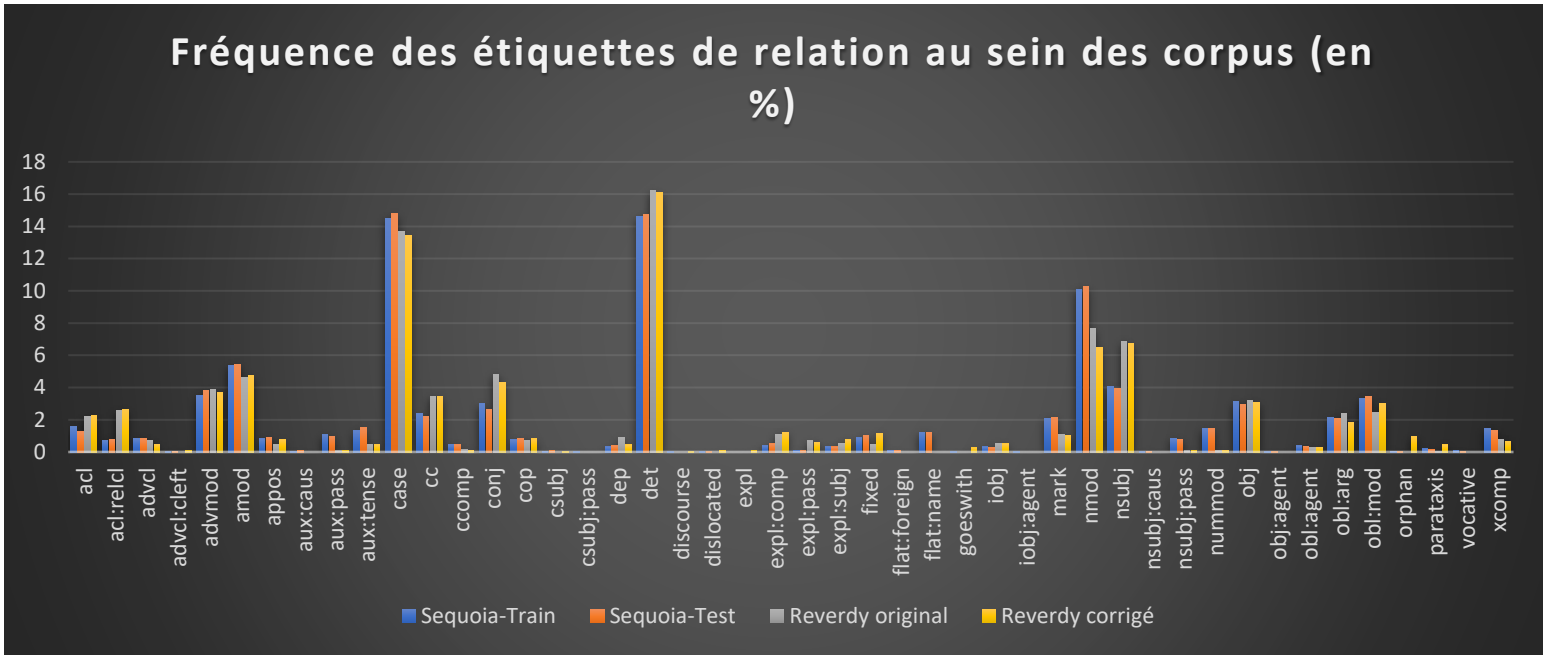
Moyenne de la longueur des dépendances des corpus

Moyenne du poids combiné (MPC)

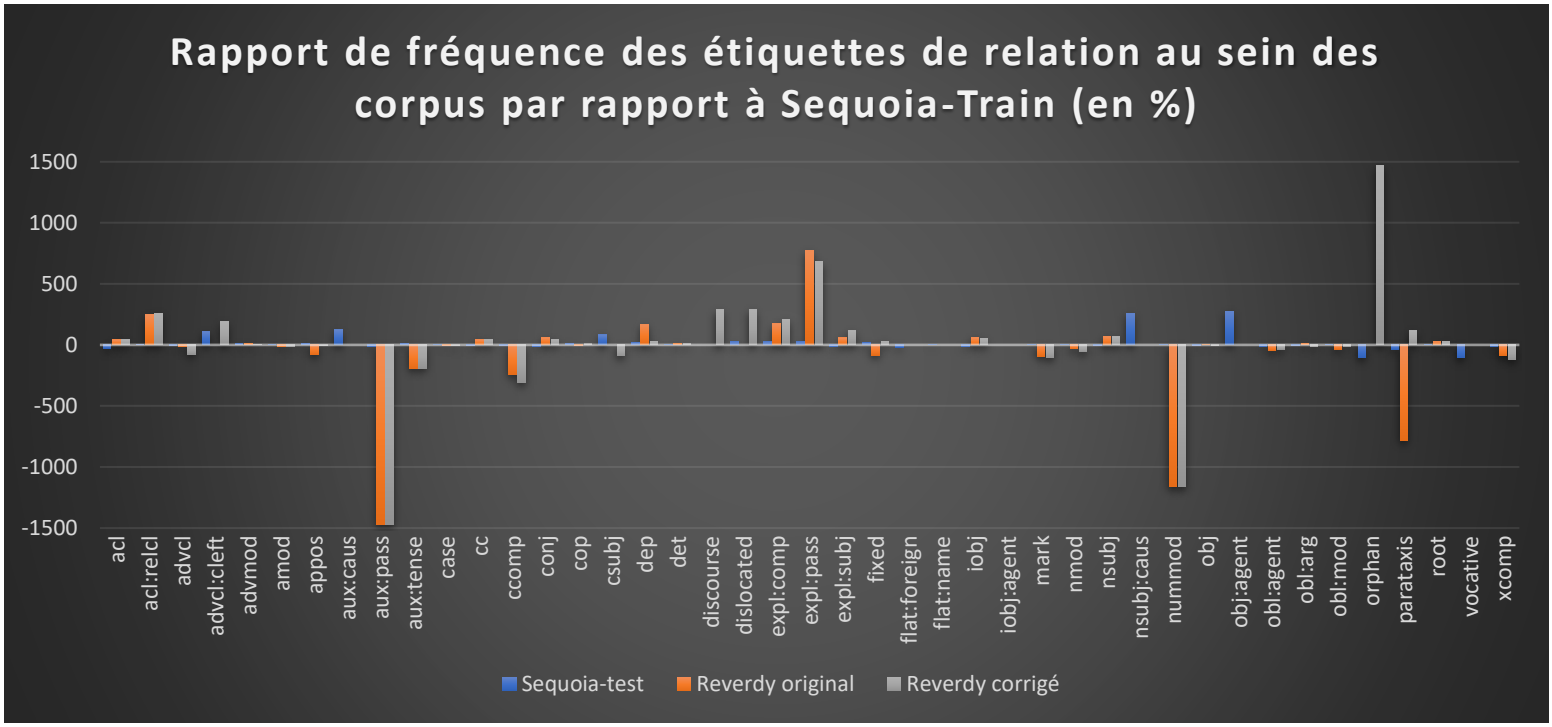
	Sequoia-Train	Sequoia-Test	Reverdy original	Reverdy corrigé segmenté	Reverdy corrigé joint
Moyenne du poids combiné	8,62	7,97	6,63	7,16	7,99

Moyenne du poids combiné des corpus

Fréquence d’apparition des étiquettes de relation



Fréquence des étiquettes de relation au sein des corpus (en %)



Rapport de fréquence des étiquettes de relation au sein des corpus
par rapport à Sequoia-Train (en %)