UFR de Langues, Littératures et Civilisations Étrangères (LLCE)

Département Sciences du Langage

Master de Sciences du Langage

Parcours Linguistique, Informatique et Technologies du Langage

Mémoire de recherche Master 1

L'alternance d'ordre des compléments du verbe *donner* en français oral spontané : focus sur le facteur *pronominalité*

Anca-Mihaela Dobrescu

Sous la direction de Cécile Fabre et Juliette Thuilier

Année universitaire 2022-2023

Remerciements

Ce travail représente le commencement d'une nouvelle étape de ma vie et je souhaite remercier un certain nombre de personnes qui m'ont soutenue, inspirée et accompagnée tout au long de cette année tout autant éprouvante que spéciale.

Je remercie en tout premier lieu mes directrices de mémoire, Madame Cécile Fabre et Madame Juliette Thuilier qui m'ont guidée, inspirée et soutenue jusqu'à la fin de ce travail. Je tiens également à remercier monsieur Ludovic Tanguy et Madame Lydia Mai Ho-Dac de leurs conseils et de toutes les connaissances transmises pendant cette année universitaire. Une autre personne importante qui m'a offert des conseils précieux a été Madame Gabriela Bîlbîie, qui coordonne l'atelier de linguistique de l'Université de Bucarest auquel je suis membre.

Mes remerciements vont également à ma famille et à mes proches qui m'ont soutenue durant tous ces mois de travail. Mention spéciale pour mon très cher ami, Ștefan Popescu, sans lequel je n'aurais pas pu arriver jusqu'à la fin de ce travail complexe.

Enfin, je remercie toute la promotion M1 LITL d'avoir embelli mon année dans un environnement rempli d'amitié et d'entraide.

Table de matières

1 Introduction....4
2 Etat de l'art....6
2.1 Contraintes catégoriques vs. contraintes préférentielles....6
2.2 Travaux effectués sur l'ordre des compléments postverbaux en français....8
2.3 Travaux similaires réalisés pour d'autres langues....13
2.3.1 L'alternance dative en anglais....13
2.3.2 L'ordre des compléments postverbaux en allemand....15
2.3.3 Les arguments des verbes ditransitifs en roumain : focus sur la pronominalité....16
2.4 La modalité orale....19
2.5 Diversité des méthodes employées....20
3 Données de l'étude....22
3.1 Corpus CEFC....22
3.2 Sous-corpus utilisés....24
3.3 Le verbe donner : les réalisations de ses arguments....25
3.4 Une première exploration des données....27
3.5 Plateforme vs. Script....29
4 Constitution du jeu de données....31
4.1 Extraction des données : sous-corpus CRFP....31
4.2 Les premières extractions....34
4.3 Évaluation de l'extraction : la méthode du Gold standard....36
4.4 Du sous-corpus CRFP au corpus CEFC....39
4.5 Améliorations réalisées et limites....42
5 Annotation des données pour l'étude de l'alternance des compléments du verbe donner....46
5.1 Annotation semi-automatique....46
5.2 Annotation automatique....49
6 Analyses statistiques de l'ordre des compléments postverbaux en français oral spontané....52
7 Discussion....60
8 Conclusions et perspectives....63
Bibliographie....64

Table des figures

Figure 1 Les étapes de la démarche expérimentale choisie……. 22
Figure 2 La répartition des phrases extraites selon le sous-corpus et la pertinence de l'annotation…….. 29
Figure 3 Chaîne de traitement pour une phrase du sous-corpus *CRFP*…… 33

Tableau 1 Les facteurs et les observés dans les travaux effectués sur l'alternance d'ordre des compléments postverbaux en français et sur des phénomènes syntaxiques similaires……. 18
Tableau 2 Les sous-corpus oraux du corpus *CEFC*……. 24
Tableau 3 Les résultats chiffrés des données de type A le sous-corpus *CFRP*……... 34
Tableau 4 Les résultats chiffrés des données de type B le sous-corpus *CFRP*……... 34
Tableau 5 Les résultats chiffrés des données de type C le sous-corpus *CFRP*……... 35
Tableau 6 La répartition des données selon les grands types de structure dans l'échantillon…. 36
Tableau 7 Résultats des mesures dans l'évaluation en fonction de grands types…… 37
Tableau 8 Résultats des mesures dans l'évaluation en fonction de chaque sous-type……... 39
Tableau 9 Les résultats chiffrés des données de type A le corpus *CEFC*……... 40
Tableau 10 Les résultats chiffrés des données de type B le corpus *CEFC*……. 41
Tableau 11 Les résultats chiffrés des données de type C dans le corpus *CEFC*…… 41
Tableau 12 La répartition des données selon le type de structure dans le corpus *CEFC*……. 42
Tableau 13 Les scores des mesures pour la couche *SN_animeite*……... 48
Tableau 14 Les scores des mesures pour la couche *SP_animeite*…… 49
Tableau 15 Les scores des mesures pour la couche *SN_accessibilite_discursive*……... 49
Tableau 16 Les scores des mesures pour la couche *SP_accessibilite_discursive*……... 49
Tableau 17 Répartition des valeurs de la variable *Longueur* dans les données…… 53
Tableau 18 La variable *Longueur* en fonction des valeurs de la variable *Ordre*…….. 53
Tableau 19 Répartition des valeurs de la variable *Complexité* dans les données……. 54
Tableau 20 La variable *Complexité* en fonction des valeurs de la variable *Ordre*……... 55
Tableau 21 Répartition des valeurs de la variable *Définitude* dans les données……... 56
Tableau 22 La variable *Définitude* en fonction des valeurs de la variable *Ordre*……. 56
Tableau 23 Répartition des valeurs de la variable *Pronominalité* dans les données……. 57
Tableau 24 La variable *Pronominalité* en fonction des valeurs de la variable *Ordre*……... 57
Tableau 25 Répartition des valeurs de la variable *Accessibilité* dans les données…… 58
Tableau 26 La variable *Accessibilité* en fonction des valeurs de la variable *Ordre*……... 58
Tableau 27 Répartition des valeurs de la variable *Animéité* dans les données……... 59
Tableau 28 La variable *Animéité* en fonction des valeurs de la variable *Ordre*……. 59

1 Introduction

La syntaxe est la partie de la linguistique qui s'occupe de la manière dont les mots se combinent entre eux pour former des phrases. La syntaxe quantitative et expérimentale est un sous-domaine de la syntaxe qui étudie la structure syntaxique des langues naturelles tout en prenant en compte la gradience des données et les différents types de contraintes qui peuvent intervenir (Bîlbîie et. al., 2021).

Elle apparaît comme un champ de recherche à la suite des difficultés de l'approche génératiste en ce qui concerne l'applicabilité de la théorie universelle selon laquelle toutes les langues du monde se conforment à une grammaire universelle (Chomsky, 1965). Dans la vision génératiste de Chomsky, la grammaticalité est une notion binaire qui s'appuie sur l'opposition grammatical/agrammatical et qui relève de la compétence d'un locuteur. Il définit la compétence comme « la connaissance que le locuteur-auditeur a de sa langue » (Chomsky, 1965, p. 4). Afin de compenser les limites du génératisme, la syntaxe quantitative et expérimentale s'appuie sur la notion d'acceptabilité qui relève de la performance. La performance suppose le fait que le locuteur fait des choix dans la langue en fonction de certains facteurs tout en n'affectant pas la grammaticalité.

Dans ce mémoire, je me propose d'étudier quels sont les facteurs, appelés plus tard contraintes préférentielles, qui interviennent dans l'ordre des mots lorsqu'un locuteur natif a le choix entre deux structures pour exprimer un contenu similaire, comme dans le cas du phénomène d'ordonnancement des compléments postverbaux en français oral. Je mènerai mes analyses sur le verbe *donner* dans le cadre de l'ordre des compléments direct et indirect. Un exemple qui illustre le phénomène syntaxique pour le verbe *donner* est le suivant :

1)

J'ai donné ce cadeau à un ami. / J'ai donné à un ami ce cadeau.

Au sens large, l'ordre des mots en français et les contraintes qui interviennent dans ce domaine ont été étudiés plus tard que pour l'anglais et à ma connaissance aucun travail sur corpus pour le français n'a été mené exclusivement pour la modalité orale, des travaux étant menés soit sur des données provenant des corpus oraux et écrits, soit seulement sur des données provenant des corpus écrits.

Dans le cadre du projet *PULCO (Prédire les Usages des LoCuteurs en français Oral)*[1], je vais réaliser mon analyse à partir du *Corpus d'Étude pour le Français Contemporain (CEFC)* qui est issu du projet *Orféo*[2] (Benzitoun et al., 2016). Comme les données du corpus ont été obtenues dans des conditions écologiques, il sera intéressant d'observer quels sont les usages des locuteurs en oral spontané, quels sont les mécanismes qui déterminent certains choix syntaxiques et quels sont les paramètres linguistiques et éventuellement extralinguistiques qui y interviennent.

A la suite des travaux élaborés dans ce domaine, les chercheurs ont pu constater l'existence de certains effets produits par des facteurs intervenant dans le phénomène d'alternance d'ordre des compléments du verbe. Cependant, pour la langue française, ils n'ont pas pu observer statistiquement l'effet de certains facteurs déjà constatés dans d'autres langues, comme la pronominalité. Contrairement à l'anglais (Bresnan et al., 2007) et à l'allemand (Kempen & Harbusch, 2004), les pronoms ne sont pas suffisamment fréquents dans la zone postverbale pour être analysés. Puisque mon étude se concentre exclusivement sur la modalité orale, la pronominalité peut s'avérer un facteur essentiel pour bien décrire les alternances des compléments, vu la haute fréquence des pronoms démonstratifs à l'oral (comme *ça*, *cela*, etc.). Il sera également intéressant d'observer la diversité des

[1] Projet coordonné par Juliette Thuilier et financé par l'*Agence Nationale de la Recherche*. (AAPG2022 PULCO JCJC)
https://anr.fr/Projet-ANR-22-CE28-0017

[2] Projet financé par l'Agence nationale de la recherche (ANR 12-CORP-0005)
https://orfeo.ortolang.fr/?locale=fr

réalisations des arguments du verbe *donner* sous le prisme de la pronominalité par la présence des pronoms clitiques dans la zone préverbale.

Ce travail de mémoire vise à tester l'importance des traits de proéminence qui ont été remarqués dans les travaux antérieurs pour le français, comme le poids grammatical, l'accessibilité discursive (Faghiri & Thuilier, 2018) et l'animéité (Thuilier et al., 2021). Le poids grammatical est envisagé de deux manières, en tant que longueur des constituants en nombre de mots, le locuteur mettant le syntagme plus court avant le syntagme plus long (distinction court avant long) et en tant que complexité syntaxique en nombre de relations de dépendance qui se créent à l'intérieur des constituants, le locuteur plaçant un syntagme plus léger avant un syntagme plus lourd (distinction léger avant lourd) (Faghiri & Thuilier, 2018). L'accessibilité discursive suppose le fait que le locuteur place d'abord les éléments déjà connus et après les éléments nouveaux (distinction *donné* avant *nouveau*) (Faghiri & Thuilier, 2018). Quant à l'animéité, le locuteur a tendance à mettre en premier un constituant dont le référent est animé plutôt qu'un constituant dont le référent est inanimé (distinction anime avant inanimé) (Thuilier, 2012b).

Puisque cette étude concerne la modalité orale, je m'attends à ce que des différences notables émergent entre les résultats des travaux antérieurs et ceux de mon étude, car selon Péry-Woodley (2000) et Bresnan et al. (2007), il semble que la modalité orale influence l'ordre des constituants. D'autres différences notables pourraient se manifester au niveau de l'effet de l'accessibilité discursive. Je fais l'hypothèse que l'effet de l'accessibilité discursive pourrait être plus évident à l'oral du fait de la tendance des locuteurs de bien distinguer l'information déjà connue de l'information nouvelle lors du discours (Péry-Woodley, 2000).

Un autre objectif de ce mémoire est de présenter la diversité des méthodes et les résultats obtenus qui ont été enregistrés autant pour le français que pour d'autres langues, comme l'anglais, l'allemand et le roumain afin de montrer la nécessité de continuer la série de travaux réalisés dans le paradigme du phénomène d'alternance d'ordre des compléments du verbe. Il est important d'analyser les similarités et les différences des résultats pour les travaux effectués car cela permet d'observer une certaine cohérence entre de différentes langues concernant les facteurs et les effets qui interviennent dans le cadre de ce phénomène syntaxique.

Compte tenu de la grande quantité de données dont je dispose (environ 4 millions de mots) et la nécessité de les extraire, de les annoter et de les analyser de point de vue statistique, j'ai utilisé des éléments spécifiques au domaine du TAL (Traitement Automatique des Langues), tels que des expressions régulières (REGEX), le langage de programmation Python, la plateforme d'annotation INCEpTION et des méthodes statistiques pour effectuer mes analyses complexes à l'aide du langage R et du logiciel Rstudio. Ce mémoire présente quelles sont les méthodes et les limites pour arriver à construire et annoter un jeu de données de haute qualité permettant l'analyse statistique du phénomène syntaxique ciblé.

Ce mémoire est structuré en huit sections. La première section représentée par cette introduction est suivie par la présentation d'un état de l'art des études antérieures effectuées sur le même phénomène syntaxique qui se retrouvent à la base de mon travail. Il s'agit des travaux effectués sur l'ordre des compléments du verbe autant en français que dans d'autres langues, comme l'anglais, l'allemand et le roumain. La troisième section vise la description des données d'étude, la sélection des sous-corpus en fonction de l'appartenance au genre ciblé, plus exactement l'oral spontané et dans un dernier temps les avantages et les inconvénients que les différentes méthodes d'extraction de données peuvent impliquer. La quatrième section décrit les modalités de constitution du jeu de données et les résultats après avoir appliqué la démarche élaborée. La cinquième section présente la démarche d'annotation des données pour l'étude de l'alternance des compléments postverbaux en français oral spontané. La sixième section présente les résultats des analyses statistiques du phénomène syntaxique étudié. La septième section est représentée par la discussion de tous les aspects observés, les limites dépassées et celles qui en demeurent encore. Dans la dernière section, je présenterai les conclusions de cette étude et des pistes qui pourraient être poursuivies dans de futurs travaux effectués dans le cadre du même phénomène syntaxique.

2 Etat de l'art

Cette section se divise en six sous-sections qui présentent les aspects théoriques de la base de l'alternance des compléments du verbe et les travaux antérieurs effectués sur ce phénomène syntaxique en français et en d'autres langues (anglais, allemand et roumain). Je commencerai par présenter l'opposition entre la notion des contraintes catégoriques et la notion des contraintes préférentielles, les dernières étant responsables du phénomène d'alternance des compléments du verbe. Dans un deuxième temps, je présenterai les études effectuées sur le français dans le cadre du même phénomène syntaxique. Puis, je présenterai les études qui ont été menées sur les trois autres langues. La sous-section suivante fera l'objet d'une illustration des effets de l'oral spontané observé au niveau de l'ordre des mots et la dernière sous-section fera un bilan de toutes les méthodes employées dans les travaux effectués dans ce domaine.

2.1 Contraintes catégoriques vs. contraintes préférentielles

L'ordre des mots peut être assez fixe ou bien libre dans les différentes langues naturelles du monde. Dans les cas où l'ordre des mots est fixe, la grammaire impose des règles strictes (contraintes catégoriques) en ce qui concerne l'emploi d'une structure plutôt que d'une autre. Les contraintes catégoriques déterminent la grammaticalité d'une phrase, ce qui conduit au fait qu'une phrase sera considérée comme agrammaticale si les normes fixées par la grammaire ne sont pas respectées.

Lorsque l'ordre des mots dans la phrase est assez libre, les locuteurs natifs choisissent une certaine structure syntaxique dans leurs discours en fonction de certains facteurs que les linguistes appellent *soft contraints*, concept traduit en français par Thuilier (2012b) par *contraintes préférentielles*. Les contraintes préférentielles exercent une influence sur l'acceptabilité des phrases et expliquent les raisons pour lesquelles le locuteur a choisi une certaine structure au détriment d'une autre, les deux structures étant possibles dans la langue (2012b).

Selon l'hypothèse de Thuilier (2012b), les contraintes préférentielles ne se retrouvent pas dans une relation d'exclusion avec les contraintes catégoriques, mais dans une relation de complémentarité étant donné que les préférences n'ont pas d'influence sur la grammaticalité. D'autre part, une contrainte peut être catégorique dans une langue et préférentielle dans une autre. L'auteure reprend comme exemples le cas de la personne grammaticale en lummi (la langue salish, Colombie Britannique) des études de Jelinek & Demers (1983, 1994) et en picuris (la langue tanoan) de l'étude Zaharlick (1982) à l'opposé de l'anglais.

Thuilier (2012b) mentionne les observations de Zaharlick (1982) sur l'utilisation des voix en anglais, plus exactement le fait que le choix est purement stylistique, tandis qu'en lummi et en picuris il existe des contraintes catégoriques qui déterminent l'emploi d'une voix plutôt que d'une autre, comme dans les exemples suivants :

2)

Agent = 1ère ou 2ème personne / Patient = 1ère ou 2ème personne

a) *ta-mo n-mia-'a n 'e -pa / *a-mo n-mia-'a n na -pa
SUJ.1SG-voir-PSF-PSÉ 2sg-par 2sg-voir-psf-psé 1SG-par
'J'ai été vu par toi' / 'Tu as été vu par moi'

b) ('e) may-mo n-'a n / (na)
2SG SUJ.2SG+OBJ.1SG-voir-PSÉ 1SG
'a -mo n-'a n
SUJ.1/3sg+obj.2SG-voir-PSÉ
'Tu m'as vu' / 'Je t'ai vu'

3)

Agent = 3ème personne / Patient = 2ème personne

a) ’a-mo n-mia-’a n sənene-pa
SUJ.2SG-voir-PSF-PSÉ homme-par
’J’ai été vu par l’homme’ (littéralement)

4)

Agent = 2ème personne / Patient = 3ème personne

a) *sənene mo n-mia-’a n ’e -pa
homme voir-psf-psé 2sg-par
’L’homme est vu par toi’

b) sənene a-mo n-’a n
homme SUJ.2SG-voir-PSÉ
’Tu as vu l’homme’

(Exemples tirés de Thuilier, 2012b qui les a repris à son tour de Zaharlick, 1982)

C’est ainsi qu’en picuris, les locuteurs n’utilisent pas la voix passive lorsque l’agent et le patient sont des arguments employés à la première ou deuxième personne et lorsqu’il y a un argument employé à la première ou deuxième personne et l’autre employé à la troisième personne, l’argument employé à la première ou à la deuxième personne sera le sujet de la phrase.

5)

Agent = 1ère personne / Patient = 3ème personne

a) x̣či-t-sən cə swəy qə
connaître-tr-1sg le homme
‘Je connais l’homme’

b) *___ (n’existe pas)
’L’homme est connu par moi’

6)

Agent = 3ème personne / Patient = 1ère personne

a) *___ (n’existe pas)
’L’homme me connaît’

b) x̣či-t-ŋ-sən ə cə swəy qə
connaître-tr-psf-1sg par le homme
’Je suis connu par l’homme’

(Exemples tirés de Thuilier, 2012b qui les a repris à son tour de Jelinek & Demers, 1983, p. 168)

Ainsi, en lummi, la voix active sera utilisée lorsque l’agent est employé à la première ou à la deuxième personne et le patient est de la troisième personne. Si l’agent est de la troisième personne et le patient est de la première ou de la deuxième personne, la voix passive sera employée.

C’est ainsi qu’en lummi et en picuris, la personne grammaticale est une contrainte catégorique dans l’alternance des voix passif/actif, tandis qu’en anglais la même contrainte est préférentielle étant donné les résultats de l’étude de Bresnan et al. (2001), selon lesquels, la contrainte de la personne grammaticale se manifeste comme une préférence au niveau statistique, comme Thuilier (2012b) l’explique.

Quant au phénomène d'alternance des compléments postverbaux en français, l'ordre est libre du point de vue syntaxique dans la majorité des cas (Abeillé et al., 2021) (Grande Grammaire du Français - GGF). Ce phénomène syntaxique est illustré par les exemples suivants :

7)

a. Paul donne [un livre]$_{SN}$ [aux enfants]$_{SP}$.

b. Paul donne [aux enfants]$_{SP}$ [un livre]$_{SN}$.

(Exemple tiré de la GGF, chapitre 3.2.1)

La GGF apporte des explications en ce qui concerne ce phénomène syntaxique. Elle précise que pour un complément nominal et pour un complément prépositionnel, il sera plus probable que le complément nominal précède le complément prépositionnel. Cependant des facteurs comme la catégorie du complément, sa longueur et la classe sémantique du verbe peuvent influencer le choix de la structure préférée par un locuteur natif. Dans la section suivante, je présenterai les études effectuées sur le français dans le cadre du phénomène de l'alternance des compléments postverbaux.

2.2 Travaux effectués sur l'ordre des compléments postverbaux en français

Blinkenberg (1928) est le premier qui parle de l'ordre des compléments postverbaux en français moderne. Il anticipe les travaux menés sur ce sujet. Tout au long de son livre, il explique que l'ordre par défaut est objet direct (OD) suivi d'objet indirect (OI). Il précise également que parfois l'ordre des compléments n'est pas contraint par la grammaire, mais il est influencé par certains facteurs comme la longueur des constituants et l'accessibilité discursive.

Une étude de référence dans le domaine de l'alternance des compléments postverbaux est représentée par l'étude expérimentale de Thuilier (2012a) qui a été menée grâce à un corpus de 956 phrases extraites du *French Treebank*, de l'*Est-Républicain* et d'*ESTER*. L'article présente une étude expérimentale qui vise à observer si certains facteurs ont un impact sur l'ordre des compléments postverbaux (compléments exprimés par un syntagme nominal - SN et un syntagme prépositionnel – SP) en français, dans les cas où il n'y a pas d'autres éléments dans la zone postverbale. Les facteurs étudiés ont été :

- la classe sémantique du verbe;
- la longueur du SN et du SP;
- le caractère défini du SN et du SP;
- la pronominalité du SN et du SP;
- le caractère animé du SN et du SP;
- le figement entre le SP et le verbe.

Comme Thuilier (2012a) l'explique, le premier qui a pris en compte le rôle du verbe pour le français a été Schmitt (1987). L'auteur a observé une tendance à suivre un ordre logique ou temporel, comme dans les exemples suivants que l'auteure a repris de lui :

- Passer de SN à SN
- Remplacer SN par SN
- Traduire de SN à SN

L'auteure s'appuie sur les travaux de Schmitt qui dit qu'il existe un ordre déterminé par la sémantique dans le cas de certains verbes comme *transposer*, *traduire* et *faire* et se propose d'élargir le champ d'application pour tous les verbes sous-catégorisant 2 compléments.

Elle étend l'hypothèse de Schmitt (1987) selon laquelle l'item verbal ainsi que sa classe sémantique influencent significativement l'ordre choisi et elle considère que le rôle du verbe n'est qu'une contrainte préférentielle.

Afin de réaliser des analyses concluantes dans le cadre de ce phénomène syntaxique, l'auteure mène son étude avec des lemmes verbaux appartenant à 14 classes sémantiques génériques extraites du dictionnaire *Les verbes du français* (LVF, Dubois & Dubois-Charlier, 1997), comme suit :

C : communication

D : don, privation

E : entrée, sortie

F : frapper, toucher

H : états physiques et comportement

L : locatif

M : mouvement sur place

N : munir, démunir

P : verbes psychologiques

R : réalisation, mise en état

S : saisir, serrer, posséder

T : transformation, changement

U : union, réunion

X : verbes auxiliaires

Thuilier (2012a) envisage l'ordre des compléments postverbaux comme un phénomène multifactoriel et considère que l'influence des différents facteurs est formalisable. Selon les précisions de l'auteure, elle propose pour ces données une modélisation statistique inspirée de Bresnan et al. (2007) et Bresnan & Ford (2010) appelée régression logistique à effets mixtes. Comme il s'agit d'une méthode de statistique inférentielle, cette méthode permet de généraliser au-delà de l'échantillon étudié. C'est ainsi qu'elle fait une évaluation de la préférence dans le cas de chaque verbe pour un ordre, indépendamment de la longueur et du caractère figé de la séquence V SP.

L'étude de Thuilier (2012a) a abouti aux résultats suivants :

- La pronominalité n'apparaît pas significative par rapport à l'ordre des compléments postverbaux contrairement à l'anglais, car en français, la cliticisation est très présente et empêche l'apparition des autres pronoms dans la zone postverbale.
- Le caractère animé des référents n'a pas d'effet significatif sur l'ordre des compléments postverbaux contrairement à l'anglais et à l'allemand.
- Le caractère défini du SN et du SP n'apparaît pas comme un facteur significatif, la proportion d'ordre SN-SP restant quasi équivalente, que le SN et le SP soient définis ou non. Ce résultat ne confirme pas l'hypothèse de Berrendonner (1987), selon laquelle l'élément en dernière position serait accompagné par un déterminant indéfini.
- En ce qui concerne la longueur, les données présentées se conforment au principe des langues SVO, selon lequel le plus long constituant a tendance à se retrouver à la fin.
- Au-delà de la longueur relative des compléments et du lien sémantique unissant le verbe et le SP, le lemme verbal ainsi que son emploi influencent l'ordre attesté. C'est ainsi qu'elle a confirmé les hypothèses de Schmitt (1987) et les a étendues à l'ensemble des verbes sous-catégorisant deux compléments postverbaux.

Une autre étude qui a analysé l'ordre des compléments postverbaux en français a été l'étude de Faghiri & Thuilier (2018). Dans le cadre de cette étude, les auteures ont effectué une expérience en production qui permet de recueillir des phrases construites par des locuteurs natifs dans des conditions contrôlées. Les auteures ont utilisé un questionnaire en ligne et à la place d'une tâche de rappel, elles ont employé une tâche de complétion, similaire à une tâche de phrase à trous. Les participants devaient compléter une phrase à partir des constituants affichés sur l'écran. Afin que les participants ne se rendent pas compte du phénomène syntaxique testé et qu'ils ne répondent pas de façon stratégique, les auteures ont ajouté un constituant formellement identique aux autres constituants. Sinon, les participants auraient pu constater qu'il s'agit d'un choix de l'ordre relatif des constituants affichés. Les auteures précisent qu'elles ont choisi 6 verbes sémantiquement comparables vu qu'ils assignent des rôles sémantiques similaires (*offrir, proposer, vendre, donner, envoyer, servir*), car elles ont voulu suivre le modèle des travaux de Schmitt (1987) et de Thuilier (2012a) qui ont montré que la sémantique du verbe influence l'ordre des compléments postverbaux.

Faghiri & Thuilier (2018) ont montré une autre dimension qui entre en jeu lors du choix de l'ordre de compléments, plus exactement la notion de poids grammatical et d'accessibilité discursive.

Le poids grammatical

Le poids peut être défini de deux manières : en tant que longueur des constituants (nombre de mots à l'intérieur d'un syntagme) ou en tant que complexité des constituants (nombre de relations de dépendance syntaxiques). Les auteures se sont demandé quelle est la manière la plus pertinente d'envisager cette contrainte qui intervient dans l'ordre des compléments du verbe. Thuilier (2012a) n'a pas montré un effet de complexité différent de l'effet de longueur en nombre de mots, mais elle a montré que la longueur en termes de nombre de syllabes a un effet moins fort que la longueur en termes de nombre de mots dans l'ordre des compléments postverbaux, comme dans l'exemple suivant :

8)

a. Jean présentera [M. Konstantin Rastapopoulos] à Marie.

b. Jean présentera à Marie [M. Konstantin Rastapopoulos].

(Exemple tiré de Thuilier, 2012b qui l'a repris de Abeillé & Godard, 2004, 2006)

L'auteure précise qu'Abeillé & Godard (2004, 2006) ont pu constater à partir de cet exemple quelle conception du poids grammatical est la plus pertinente dans une analyse. Plus exactement, l'ordonnancement des compléments en français serait plus susceptible d'être influencé par la complexité syntaxique en termes de mots que par la longueur syllabique des constituants. Pour donner suite à ces observations, les auteures de cette étude qui date de 2018 se sont demandé si le poids est associé exclusivement à la complexité des constituants en termes de mots ou bien s'il faut prendre toujours en compte le nombre de syllabes des constituants.

Après avoir effectué l'étude en variant dans leurs items la longueur en tant que nombre de constituants, mais en gardant le même nombre de syllabes, les auteures ont observé que le poids grammatical ne peut pas être associé exclusivement à la complexité syntaxique, mais aussi à la longueur en termes de constituants. Selon les auteures, la complexité syntaxique a plus de chances que la longueur de bouleverser l'ordre SN-SP.

L'accessibilité discursive

L'accessibilité discursive est définie comme le caractère donné ou nouveau des référents et a été mentionnée comme un facteur pertinent dans certains travaux sur le français comme Blinkenberg (1928), mais aucune étude empirique n'a permis de montrer son influence dans cette question syntaxique. Des notions corrélées à l'accessibilité discursive sont le thème et le rhème, comme précisé

par Péry-Woodley (2000). C'est ainsi que le thème est envisagé comme la partie initiale de la phrase qui décrit sur quoi porte la proposition en ayant un référent déjà connu tandis que le rhème représente la partie finale de la phrase et la plus saillante de l'information nouvelle (indépendamment du statut du référent). Par conséquent, il existe une tendance à mettre les éléments déjà connus en début de phrase tandis que les éléments nouveaux sont gardés pour la fin de la phrase, comme dans les exemples suivants :

9)

a. A qui Moscou a-t-il envoyé un message d'appui ?

b. Moscou a envoyé un message d'appui au mouvement des cent-un.

10)

a. Qu'est-ce que Moscou a envoyé au mouvement des cent-un ?

b. Moscou a envoyé au mouvement des cent-un un message d'appui.

(Exemples tirés de Faghiri & Thuilier, 2018 qui ont été repris de Berrendonner, 1987)

Ainsi, le locuteur aura tendance à placer la réponse de la question (l'élément nouveau) à la fin de la phrase.

Faghiri & Thuilier (2018) ont fait l'hypothèse qu'il y ait une préférence pour le donné par rapport au nouveau et que l'effet d'accessibilité devrait être plus important si le poids du SN et du SP est similaire. Afin d'observer l'effet de l'accessibilité discursive, les auteures ont ajouté une phrase avant la phrase ciblée pour fixer le contexte. Quand il s'agissait d'un SP donné, le référent était présent dans la phrase qui fixait le contexte, alors que dans le cas d'un SP nouveau, le référent était absent dans la phrase-contexte.

Les résultats de cette étude ont montré que même s'il y a une préférence générale pour l'ordre SN-SP, dans le cas d'un SN est plus lourd que le SP, il existe plus de chances que le bouleversement de l'ordre se produise. Quant à l'aspect discursif, le fait d'avoir un SP donné dans la phrase augmente les chances d'avoir un ordre SP-SN, mais cet effet peut disparaître lorsque le SN est court. Même si l'effet du poids s'est avéré très fort par rapport à l'effet de l'accessibilité discursive, les choix des locuteurs ont été partiellement expliqués par la différence du statut nouveau ou donné du réfèrent. Comme dans leur étude l'accessibilité des référents ne se retrouve pas en lien avec la pronominalité, l'effet du statut donné ou nouveau des référents ne se retrouve pas en lien avec la taille réduite des pronoms.

Comme les études effectuées sur le français auparavant n'ont pas pu montrer l'effet de l'animéité de façon directe ou de façon indirecte dans l'ordre des compléments, Thuilier et al. (2014) ont effectué une étude afin d'essayer de trouver une certaine cohérence du français avec l'anglais et l'allemand en ce qui concerne l'effet d'animéité qui intervient.

L'animéité

L'animéité est une propriété sémantique des référents qui se manifeste aux différents degrés : les plus animés sont les humains et les moins animés et moins concrets sont les entités abstraites. Thuilier et al. (2014) ont essayé d'observer l'effet de l'animéité indépendamment de l'effet du poids.

Cette étude s'est basée sur une analyse du corpus *French Treebank* dont les données proviennent du journal *Le Monde*, le corpus *Est Républicain* et sur deux corpus parlés le corpus radiophonique public *ESTER,* la partie française de *C-ORAL-ROM*) et sur deux questionnaires psycholinguistiques pour

collecter les jugements d'acceptabilité des locuteurs natifs français. Ils ont également repris des verbes proposés par Schmitt (1987).

Comme les auteurs ont observé que le poids exerce un effet considérable dans l'analyse des corpus, ils ont voulu le neutraliser pour la première expérience psycholinguistique afin d'observer si les autres facteurs exercent une influence sur l'ordre des compléments du verbe. C'est pour cette raison qu'ils ont créé une tâche de jugement d'acceptabilité, dont les conditions permettent de contrôler l'interférence des autres facteurs sur l'effet recherché. Ils ont extrait 23 phrases des corpus étudiés avec des compléments de longueur égale, comme dans l'exemple suivant :

11)

Pierre fonce dans la nuit porter [la bonne nouvelle] [à sa fiancée]. (ER)

(Exemple tiré de Thuilier et al., 2014)

Pour cette tâche, ils ont testé 25 participants (étudiants de l'Université Paris Diderot) sur l'ordre des compléments postverbaux en français en utilisant une échelle de 1 à 5 points sur des items qui contenaient des distracteurs. Les résultats de cette expérience n'ont pas montré l'existence d'un effet de l'animéité sur l'ordre des compléments postverbaux. Afin d'observer l'effet de l'animéité, les auteurs de l'article ont réalisé une deuxième expérience dont le design expérimental neutralise l'effet d'autres facteurs en dehors de la variable *ordre* et de la variable *animéité*. C'est ainsi qu'ils ont mis en place un questionnaire qui a été complété par 38 participants étudiants à Paris. Les 16 items pouvaient être évalués sur une échelle de 1 à 5, les constituants avaient la même longueur et étaient définis, l'OD étant toujours inanimé, comme dans l'exemple :

12)

Il faut que les Israéliens maintenant, dans les prochaines semaines, dans les prochains mois

a. [DO-IO/-anim] donnent les réponses précises à ces questions.

b. [IO-DO/-anim] donnent à ces questions les réponses précises.

(Exemple tiré de Thuilier et al., 2021 qui a été repris de Thuilier et al., 2014)

En regardant les résultats, les auteurs n'ont pas pu mettre en évidence d'effet significatif de l'animéité sur l'ordre des compléments postverbaux en français malgré le fait qu'il y avait une légère préférence de l'ordre animé avant inanimé et défini avant indéfini (observé aussi pour l'allemand et l'anglais).

Une autre étude qui s'est proposée d'observer ce facteur a été Thuilier et al. (2021). Les auteurs ont analysé si l'animéité a un impact sur l'ordre linéaire ou sur l'attribution de fonctions grammaticales. Selon les auteurs de cette étude, le fait que cet effet a été prouvé pour le français dans l'alternance des voix semble soutenir l'hypothèse selon laquelle, il existe une influence indirecte de l'animéité.

Thuilier et al. (2021) proposent une tâche de rappel qui pourrait prouver que l'animéité est un trait de proéminence dans l'ordre linéaire des compléments ou sur l'affectation des fonctions grammaticales. Les auteurs de cette étude ont voulu tester l'alternance des voix passif/actif, l'alternance des compléments postverbaux et l'alternance des syntagmes nominaux conjoints coordonnés.

Les verbes choisis pour les items appartiennent à deux classes sémantiques (verbes de transfert et verbes de communication) qui forment une classe homogène grâce à leur structure (Les deux déterminent l'apparition de la préposition *à* et assignent des rôles similaires à leurs arguments). Les participants ont été répartis en deux groupes expérimentaux : Le groupe 1, trente-neuf participants, a vu les items avec alternance de voix ; et le groupe 2, trente-huit participants, a été exposé à la coordination et aux verbes ditransitifs. La phase d'étude a été visuelle et la phase de rappel a été orale. Les participants ont été testés individuellement dans une pièce calme.

Les résultats de Thuilier et al. (2021) ont été surprenants : au lieu de pouvoir observer un effet d'animéité, comme les auteurs s'y attendaient au début, ils ont pu voir plutôt un effet d'anti-animéité qui suppose la préférence d'un ordre objet direct inanimé-objet indirect animé, comme dans les exemples suivant :

13)

a) Le chef du projet a confié un nouveau budget à un décorateur. [OD inanimé, OD-OI]
b) Le chef du projet a confié à un décorateur un nouveau budget. [OD inanimé, OI-OD]
c) Le chef du projet a confié un agent commercial à un décorateur. [OD animé, OD-OI]
d) Le chef du projet a confié à un décorateur un agent commercial. [OD animé, OI-OD]

(Exemple tiré de Thuilier et al., 2021)

Les locuteurs ont eu tendance à inverser l'ordre OI OD vers OD OI quand l'OD est inanimé, donc de préférer la phrase a) au détriment de la phrase b).

En résumé, les études présentées effectuées sur le français dans le cadre de ce phénomène syntaxique ont identifié les facteurs suivants qui exercent une influence sur l'ordre de compléments postverbaux :

- le lemme verbal et la classe sémantique du verbe ;
- le lien sémantique unissant le verbe et le SP ;
- le poids grammatical (en tant que longueur de constituants, ce qui donnera un effet de court avant long et en tant que complexité syntaxique, ce qui donnera un effet de léger avant lourd) ;
- l'accessibilité discursive (avec un effet de donné avant nouveau) ;
- l'animéité (avec un effet d'OD inanimé avant OI animé).

2.3 Travaux similaires réalisés pour d'autres langues

Cette section présente trois études effectuées sur le phénomène syntaxique d'alternance d'ordre des compléments du verbe dans le cadre des autres langues. Je commencerai par une étude effectuée sur l'anglais (section 2.3.1), puis je continuerai avec une étude réalisée pour l'allemand (section 2.3.2) et enfin une étude menée sur le roumain (section 2.3.3).

2.3.1 L'alternance dative en anglais

En anglais, le phénomène syntaxique similaire au phénomène présent en français s'appelle l'alternance dative. Dans le cadre de ce phénomène, les constructions dont l'ordre varie sont la construction à double objet et la construction à SP-datif. L'alternance dative en anglais apparaît très bien illustrée dans l'exemple suivant :

14)

a. Susan gave [toys] [to the children].

'Susan donne le jouet aux enfants.'

b. Susan gave [the children] [toys].

'Susan donne aux enfants le jouet.'

(Exemple extrait de Bresnan et al., 2007)

L'étude de référence de Bresnan et al. (2007) présente une étude expérimentale qui vise à montrer qu'il est possible de mesurer l'impact de plusieurs variables en même temps dans le cadre d'une analyse sur corpus. L'étude se concentre sur les contraintes préférentielles qui régissent le choix entre

une construction à double objet et une construction à SP-datif dans le cadre de l'alternance dative manifestée en anglais en tant que phénomène syntaxique. Dans le cadre de cette étude, les auteurs ont pris en considération 14 variables :

- la classe sémantique du verbe;
- l'accessibilité du destinataire;
- l'accessibilité du thème;
- la pronominalité du destinataire;
- la personne grammaticale du destinataire;
- le nombre du destinataire;
- la pronominalité du thème;
- la définitude du destinataire;
- la définitude du thème;
- l'animéité du destinataire;
- le caractère concret du thème;
- le nombre du thème;
- le parallélisme structural dans le dialogue;
- la longueur relative du destinataire et du thème.

Bresnan et al. (2007) ont mené une étude expérimentale grâce à un corpus de 3 millions de mots réalisé à partir des conversations téléphoniques spontanées. Ils utilisent 2360 de constructions datives annotées. Les auteurs cherchent à savoir quels facteurs déterminent le choix de construction, autrement dit quelles sont les contraintes préférentielles qui interviennent lors de l'activité discursive d'un natif. Les anciennes études sur l'alternance dative utilisaient de très petits échantillons, ce qui limitait l'étude simultanée de plusieurs variables, c'est-à-dire facteurs qui déterminent la préférence d'une construction au détriment d'une autre. Afin de prédire le choix d'une structure, les auteurs de cette étude ont construit trois modèles statistiques en utilisant la méthode de régression logistique. Cette méthode permet de prédire le comportement d'une variable binaire à partir de plusieurs variables prédictrices.

Les principaux résultats de cette étude sont :

- Le caractère donné est plus proéminent que le caractère non-donné, c'est-a-dire l'élément déjà connu se trouve dans la phrase avant l'élément nouveau de point de vue informationnel.
- Le fait qu'il s'agit d'un pronom est plus proéminent qu'un non-pronom, c'est-à-dire un syntagme dont la tête est un pronom se retrouve avant un syntagme dont la tête n'est pas un pronom.
- Le caractère animé est plus proéminent que le caractère non-animé, c'est-à-dire un syntagme dont le référent est animé se trouvera dans la phrase avant un syntagme dont le référent est inanimé.
- Le caractère défini est plus proéminent que le caractère non-défini, ce qui veut dire qu'un syntagme dont le référent est défini se trouvera dans la phrase avant un syntagme dont le référent est indéfini.
- L'ordre court avant long sera privilégié, ce qui veut dire qu'un syntagme plus court se trouvera dans la phrase avant un syntagme plus long.
- Si le constituant qui présente le plus de propriétés proéminentes est le destinataire, la construction à double objet sera privilégiée et si ce constituant est le thème, la construction à SP datif sera privilégiée.

En ce qui concerne le phénomène correspondant à l'alternance dative pour le français, notamment l'ordre des compléments postverbaux (qui ne suppose pas deux constructions syntaxiques différentes comme en anglais, mais représente exclusivement une question d'ordre), les auteurs ont essayé de répliquer ce qui a été prouvé pour l'anglais et éventuellement arriver aux mêmes résultats, ce qui donnerait une certaine cohérence entre les langues du monde. A la suite des études effectuées pour

l'anglais, Thuilier (2012a) s'est inspirée de Bresnan et al. (2007) et de Bresnan & Ford (2010) pour voir si elle arrive aux mêmes résultats en français.

Cette étude a été une source d'inspiration pour les travaux qui ont été faits pour le français et reste une étude de référence dans le domaine de la syntaxe quantitative et expérimentale.

2.3.2 L'ordre des compléments postverbaux en allemand

Kempen & Harbush (2004) ont constaté comme leurs prédécesseurs que la grammaire de l'allemand n'impose pas de contraintes strictes sur l'ordre linéaire du sujet, de l'objet indirect et de l'objet direct et ils ont voulu observer quelles sont les contraintes préférentielles qui interviennent pour un natif allemand lorsqu'il fait un choix de construction d'une part dans son discours à l'oral et d'autre part à l'écrit. Ils ont effectué une étude sur un corpus dont les données ont été extraites à partir de la base de données *NEGRA-II* contenant environ 20 000 phrases de journaux annotées en détail et à partir des annotations faites à la main pour l'animéité et la définitude. Ils ont classifié leurs phrases extraites en fonction des deux catégories : les phrases intransitives (les phrases qui contiennent un sujet et un complément d'objet indirect) et les phrases transitives qu'ils ont réparties en deux sous-catégories, les monotransitives (les phrases qui contiennent un sujet et un complément d'objet direct) et les ditransitives (les phrases qui contiennent un sujet, un complément d'objet direct et un complément d'objet indirect).

Le trait de proéminence de la pronominalité dans le cadre du phénomène de l'ordre des compléments du verbe en allemand est visible dans les exemples suivants :

15)

Pronom datif

Der	alte	Mann	hat	ihm	das	Buch	geschenkt.
Le	vieux	homme	a	3SG.DAT.	le	livre	offert

'Le vieux homme lui a offert le livre.'

16)

Pronom accusatif

Der	alte	Mann	hat	es	seinem	Sohn	geschenkt.
Le	vieux	homme	a	3SG.ACC.	son-DAT	fils	offert

'Le vieux homme l'a offert à son fils.'

(Exemples tirés de Thuilier, 2012b)

C'est ainsi que les pronoms personnels dans les cas accusatif (représentant un OD) et datif (représentant un OI) ont tendance à être placés avant lorsqu'ils apparaissent à côté des groupes nominaux complémentaires dans l'alternance des compléments postverbaux. Thuilier (2012b) explique par rapport à cette étude effectuée sur l'allemand qu'il existe plusieurs facteurs qui favoriseraient l'ordre pronominal - non-pronominal, comme :

- le nombre réduit de syllabes (les pronoms sont soit monosyllabiques, soit bisyllabiques)
- les constituants à tête pronominale qui ne sont pas complexes
- les constituants à tête pronominale accessibles de point de vue discursif (ils se retrouvent soit en emploi anaphorique, soit déictique).

Cependant, Bresnan et al. (2007) ont démontré pour l'anglais que l'effet de la pronominalité est indépendant de l'effet du poids et de l'accessibilité discursive.

L'animéité a été définie comme « se référant à un humain ou un animal, ou un collectif d'humains/animaux ». En cas de doute sur un référent, ils ont considéré le référent en question comme animé. En concordance avec la tendance générale de mettre les pronoms avant les non-pronoms, l'ordre OD sujet est plus fréquent lorsque le sujet non-pronom est inanimé (85%). Quand le sujet est animé, la tendance de choisir cet ordre est moins fréquente (51%), ce qui conduit à l'idée que l'animéité joue un rôle si important que les locuteurs ont tendance à mettre les référents animés en premier. Pour ce qui est de l'ordre OI sujet, les locuteurs ont mis le sujet en premier lorsque l'OI était inanimé (93%). Quand l'OI était animé, ils ont mis le sujet en premier moins souvent (54%). D'après leurs résultats, il semble que l'animéité affecte non seulement la structure fonctionnelle de la phrase, mais aussi sa structure linéaire.

Étant donné ces résultats, il semble que le français ne se conforme pas au modèle de l'anglais et de l'allemand en ce qui concerne l'effet d'animéité et de pronominalité. Contrairement aux études effectuées sur l'anglais et l'allemand, Thuilier et al. (2021) ont observé un effet d'anti-animéité chez les locuteurs français qui ont tendance à mettre en avant plutôt les compléments d'objet direct inanimés que les compléments d'objet indirect animés. Pour ce qui est de l'effet de pronominalité en français, Thuilier (2012b) fait l'hypothèse que l'effet de pronominalité (généré par les compléments qui sont exprimés par un pronom personnel défini) en français est réduit contrairement à l'allemand dû à la présence des pronoms clitiques dans la zone préverbale. C'est-à-dire, la position postverbale des compléments réduit la possibilité de retrouver des compléments pronoms personnels dans la zone postverbale, les pronoms se situant d'habitude avant le verbe.

2.3.3 Les arguments des verbes ditransitifs en roumain : focus sur la pronominalité

Le chapitre 6 de Tigău (2020) présente une étude expérimentale sur la configuration ditransitive en roumain en prenant comme point de repère l'alternance dative en anglais. Cette étude décrit quel est le rôle joué par les clitiques dans le cas des verbes ditransitifs. Dans cette étude, il s'agit de 3 expériences qui ont pour but de vérifier les dépendances qui s'établissent entre les 2 arguments internes de la construction ditransitive. Les trois expériences ont été similaires dans la conception et ont vérifié l'acceptabilité des phrases dans lesquelles l'auteure a fait varier : a) l'ordre des mots en surface ; b) la direction du liage (le fait que l'OD soit lié à l'OI ou l'OI a l'OD de point de vue de l'interprétation référentielle des éléments) ; c) la présence d'un clitique datif doublant l'OI. Ce qui différencie les trois expériences a été le type d'OD employé (s'il a un marquage différentiel de l'objet ou s'il n'en a pas - MDO) .

L'étude de Tigău (2020) propose une tâche de jugements d'acceptabilité chez les locuteurs natifs roumains réalisée par l'intermédiaire d'un questionnaire. Elle a fait varier la présence d'un clitique datif doublant l'OI et la présence du marquage différentiel de l'objet (MDO) réalisé par l'intermédiaire de la préposition *pe* du roumain.

OD[-marqué] + OI[-doublé]

a.	Editorii	au	trimis	fiecare	carte	autorului	ei
	Éditeurs.les	ont	envoyé	chaque	livre	auteur.DAT	son
	pentru	corecturile	finale.				
	pour	corrections.les	finales				

'Les éditeurs ont envoyé chaque livre à son auteur.'

OD[-marqué] + OI[+doublé]

b.	Editorii	i-au	trimis	fiecare	carte	autorului	ei
	Éditeurs.les	CL.DAT.SG.ont	envoyé	chaque	livre	auteur.DAT	son

pentru corecturile finale.

pour corrections.les finales

'Les éditeurs ont envoyé chaque livre à son auteur.'

OD[+marqué] + OI[-doublé]

c. Comisia a repartizat pe fiecare medic rezident

Comité.le a assigné MDO chaque médecin résident

unor foşti profesori de-ai lui.

certains.DAT anciens professeurs.GEN.SG

'Le comité a assigné chaque médecin résident à l'un de ses anciens professeurs.'

OD[+marqué] + OI[+doublé]

d. *Comisia le-a repartizat pe fiecare medic

Comité.le CL.DAT.PL-a assigné DOM chaque médecin

rezident unor foşti profesori de-ai lui.

résident certains.DAT anciens professeurs.GEN.SG

'Le comité a assigné chaque médecin résident à l'un de ses anciens professeurs.'

OD[+doublé][+marqué] + OI[-doublé]

e. Comisia l-a repartizat pe fiecare medic

Comité.le CL.ACC.SG.M-a assigné DOM chaque médecin

rezident unor foşti profesori de-ai lui.

résident certains.DAT anciens professeurs.GEN.SG

'Le comité a assigné chaque médecin résident à l'un de ses anciens professeurs.'

OD[+doublé][+marqué] + OI[+doublé]

f. Comisia li l-a repartizat pe

Comité.le CL.DAT.PL CL.ACC.SG.M-a assigné DOM

fiecare medic rezident unor foşti profesori de-ai lui.

chaque médecin résident certains.DAT anciens professeurs.GEN.SG

'Le comité a assigné chaque médecin résident à l'un de ses anciens professeurs.'

(Exemples tirés de Tigău, 2020)

Selon les résultats, les phrases avec un OI doublé par un clitique ont été considérées moins acceptables que celles qui n'avaient pas l'OI doublé par un clitique. Ainsi, les phrases 19-a, 19-c et 19-e ont été considérées comme plus acceptables que les phrases 19-b, 19-d et 19-f étant donné l'absence du redoublement par des pronoms clitiques dans la zone postverbale.

A travers ses résultats, l'auteure a démontré que le roumain dépend plutôt de la classe sémantique du verbe que de la présence ou l'absence des clitiques. Pourtant, les phrases avec un OI doublé par un

clitique ont été considérées moins acceptables que celles qui n'avaient pas le OI doublé par un clitique.

Une autre observation faite par l'auteure a été que l'ordre de surface des mots des deux arguments internes a un impact significatif sur les jugements d'acceptabilité. L'auteure a également remarqué que l'OD aura priorité devant l'OI et qu'il est possible que la personne grammaticale ait des influences sur l'ordre choisi des compléments postverbaux chez les locuteurs natifs. La priorité peut, cependant, changer en fonction de la spécification des caractéristiques des deux objets.

L'étude expérimentale de Tigău (2020) montre une autre dimension du phénomène de l'alternance des compléments du verbe, notamment les changements qui sont générés au niveau de l'acceptabilité des phrases une fois l'OI doublé par un pronom clitique dans la zone préverbale.

Pour résumer, les facteurs et leurs effets qui ont été observés jusqu'à présent dans les travaux effectués sur l'alternance des compléments postverbaux en français et sur des phénomènes syntaxiques similaires en anglais, allemand et roumain sont présents dans le tableau suivant :

Facteurs	**Français**	**Anglais**	**Allemand**	**Roumain**
Lemme verbal	Effet observé	Effet observé	Pas observé	Effet observé
Classe sémantique du verbe	Effet observé	Effet observé	Pas observé	Effet observé
Poids grammatical	Effet de court-long et léger-lourd	Effet de court-long et léger-lourd	Pas observé	Pas observé
Accessibilité discursive	Effet de donné-nouveau	Effet de donné-nouveau	Pas observé	Pas observé
Animéité	Effet d'anti-animéité	Effet d'animé-inanimé	Effet d'animé-inanimé	Pas observé
Pronominalité	Pas observé	Effet de pronom-non-pronom	Effet de pronom-non-pronom	Effet des clitiques sur l'acceptabilité
Définitude	Absence d'effet	Effet défini-non-défini	Effet défini-indéfini	Pas observé
Longueur	Effet de court-long	Effet court-long	Effet court-long	Pas observé
Rôle sémantique	Effet en lien avec l'animéité	Effet observé	Effet observé	Pas observé
Affectation de fonctions grammaticales	Effet en lien avec l'animéité	Pas observé	Effet observé	Pas observé

Tableau 1 *Les facteurs et les observés dans les travaux effectués sur l'alternance d'ordre des compléments postverbaux en français et sur des phénomènes syntaxiques similaires*

La section suivante illustre quelles sont les spécificités et les changements qui pourraient intervenir pour la modalité orale dans le cadre de l'analyse de l'ordre des mots dans la phrase.

2.4 La modalité orale

La modalité orale présente des spécificités par rapport à la modalité écrite et peut influencer considérablement l'analyse de l'ordre des arguments du verbe et les facteurs qui interviennent dans le cadre du phénomène syntaxique concerné.

Wang et al. (2020) mentionnent les spécificités qui peuvent apparaître dans le cas d'une modalité orale au niveau du corpus. Ces spécificités d'un corpus oral, comme le corpus *ODIL_Syntax* sont représentées par l'existence de certaines disfluences qui interviennent dans le cadre d'un discours. Les disfluences que Wang et ses collaborateurs mentionnent sont :

- les inachèvements (peuvent être définis comme des interruptions pour poursuivre en débutant un nouvel énoncé qui ne partage aucun lien syntaxique avec le début de son tour de parole);
- les entassements paradigmatiques (peuvent prendre la forme des répétitions, des auto-corrections ou des reprises).

Da Cunha (2021) a pu relever que selon la modalité de transmission des données (écrite ou orale) des configurations différentes s'établissent au niveau de la langue française pour l'alternance actif/passif. Les configurations reposent sur un type d'alternance de réalisation syntaxique d'arguments qui gardent les mêmes rôles sémantiques, comme Da Cunha (2021) l'explique en s'appuyant sur Creissels (2006) et Haspelmath (2020). Da Cunha (2021) a travaillé sur un corpus écrit, notamment le *French Treebank (FTB)* et trois corpus oraux appartenant au *Corpus d'Étude pour le Français Contemporain (CEFC)*. Selon l'auteur, l'oral a tendance à suivre un certain codage harmonique sujet pronominal et défini, objet nominal et indéfini, tandis que l'écrit présente plutôt un codage non harmonique (par exemple sujets nominaux ou indéfinis). Vu ces aspects, il a observé que la contrainte de définitude est très forte à l'oral.

Une étude qui a montré comment la modalité orale peut changer l'ordre des compléments du verbe dans le cadre de l'alternance dative a été l'étude de Bresnan (2007). L'auteure a montré que le changement de modalité (écrite et orale) peut influencer les résultats. C'est ainsi qu'elle a constaté pour les données dont elle disposait qu'à la différence de l'anglais oral, il est légèrement plus probable d'utiliser une structure prépositionnelle à l'écrit qu'à l'oral.

Dans le cadre de l'analyse de l'alternance d'ordre compléments du verbe en français, l'ordre peut être influencé aussi à cause de la modalité orale et une étude qui mentionne le rôle que pourrait avoir la modalité orale dans l'ordre choisi des constituants est l'étude de Péry-Woodley (2000). Elle décrit dans son ouvrage la vision des linguistes en ce qui concerne l'ordre des mots qui serait « la résultante de forces diverses, parmi lesquelles, outre la syntaxe, la structure sémantique (Firbas, 1972), l'iconicité expérientielle, les contraintes liées au traitement en temps réel dans l'oral spontané (Enkvist, 1985) » (Péry-Woodley, 2000, p. 20). C'est ainsi que de différences notables pourraient se manifester entre l'ordre des constituants dans le cas d'une modalité écrite et dans le cas d'une modalité orale.

En résumé, la modalité orale peut apporter des éléments nouveaux dans l'analyse d'un phénomène syntaxique qui vise l'ordre des mots dans la phrase.

La section suivante fait une synthèse sur les méthodes employées dans les études déjà présentées dans cet état de l'art en soulignant l'importance des méthodes possibles dans l'analyse du phénomène syntaxique d'alternance des compléments du verbe.

2.5 Diversité des méthodes employées

La syntaxe quantitative et expérimentale propose une diversité des méthodes qui contribuent aux analyses et à l'interprétation des résultats. Il existe deux approches principales que les linguistes choisissent dans le cadre de leurs études : l'analyse du phénomène syntaxique ciblé sur corpus et l'analyse du même phénomène par l'intermédiaire des questionnaires psycholinguistiques.

Selon Gilquin & Gries (2009), les études menées sur corpus présentent de nombreux avantages par rapport aux études psycholinguistiques, comme :

- Les données proviennent de contextes naturels et elles permettent donc d'étudier les questions de registre/genre qui sont difficiles à étudier de manière expérimentale.
- Il est possible d'étudier un plus grand nombre de données que dans le cas des études psycholinguistiques.
- La diversité ou le bruit des données (ce qui ne fait pas l'objet de l'étude parmi les données à disposition) peut aujourd'hui être mieux gérée avec des statistiques multifactorielles.

Parmi les études mises en discussion dans le cadre du phénomène de l'alternance des compléments du verbe, celles qui utilisent cette première approche sont Thuilier (2012a, b), Bresnan et al. (2007), Thuilier et al. (2014) et Da Cunha (2021) et Kempen & Harbusch (2004). Afin d'estimer l'impact des facteurs étudiés de manière isolée, mais aussi en interaction avec d'autres facteurs, les auteurs de ces études ont réalisé des analyses statistiques.

L'autre approche, celle qui utilise des questionnaires psycholinguistiques, suppose des tâches qualitatives, comme la tâche de choix forcé et la tâche en oui ou non et des tâches quantitatives comme la tâche de jugement avec échelle et la tâche de jugement par estimation de grandeur. Dans la tâche de choix forcé, le participant doit choisir la version préférée dans la conscience du phénomène syntaxique étudié. Dans la tâche de jugement avec échelle, le locuteur doit accorder aux phrases vues des points sur une échelle prédéfinie. Cette méthode a pour avantage que le participant n'a pas accès directement à la question de recherche et il peut manifester ses intuitions qui sont souvent graduelles. Une autre tâche présente dans les études est la tâche de rappel qui suppose souvent une complétion des phrases. De manière générale, une tâche de rappel suppose 3 étapes :

- Les participants entendent des phrases.
- Les participants sont soumis à une tâche de distraction pour que les participants ne retiennent pas par cœur les phrases entendues.
- Les participants doivent reproduire les phrases entendues préalablement.

D'après Gilquin & Gries (2009), les études expérimentales réalisées par l'intermédiaire des questionnaires psycholinguistiques ont aussi leurs avantages par rapport aux études réalisées sur corpus, comme :

- Elles permettent d'étudier les phénomènes rares qui ne se retrouvent pas facilement dans les corpus.
- Elles permettent de contrôler systématiquement toutes les variables qui entrent en jeu dans le cadre du phénomène syntaxique étudié.
- Selon la nature de l'expérience, elles permettent d'effectuer des tâches en ligne.

Parmi les études mises en discussion dans le cadre du phénomène de l'alternance des compléments du verbe, celles qui utilisent cette deuxième approche sont : Thuilier et al. (2014), Faghiri & Thuilier (2018), Thuilier et al. (2021) et Tigău (2020). Thuilier et al. (2014) et Tigău (2020) proposent des questionnaires pour collecter les jugements d'acceptabilité auprès de locuteurs en utilisant une échelle de 1 à 5, tandis que Faghiri & Thuilier (2018) et Thuilier et al. (2021) proposent des tâches de production, la première étant une tâche de complétion de phrases et la deuxième étant une tâche de rappel de phrases. Les auteurs de ces études ont élaboré des protocoles expérimentaux qui facilitent l'observation de l'effet des facteurs étudiés de manière isolée et en interaction avec d'autres facteurs.

Gilquin & Gries (2009) considèrent que la meilleure approche serait d'utiliser les deux méthodes afin de compenser les limites de chaque approche, ce qui permettra des analyses pertinentes dans le cadre du phénomène syntaxique étudié.

Dans ce mémoire, je vais me concentrer sur la première approche. Plus exactement j'effectue une étude exclusivement sur corpus dans le cadre de l'analyse de la réalisation des arguments du verbe *donner* et de l'analyse de l'ordre des compléments postverbaux pour le même verbe. Les sections suivantes (section 3, section 4, section 5 et section 6) présentent les données utilisées et la méthodologie mise en place dans cette étude.

3 Données de l'étude

La démarche expérimentale que j'ai choisie comprend plusieurs étapes qui permettent l'analyse de la réalisation des arguments du verbe *donner* et l'analyse d'ordre des compléments postverbaux en français oral spontané. Ces étapes sont synthétisées dans le schéma suivant :

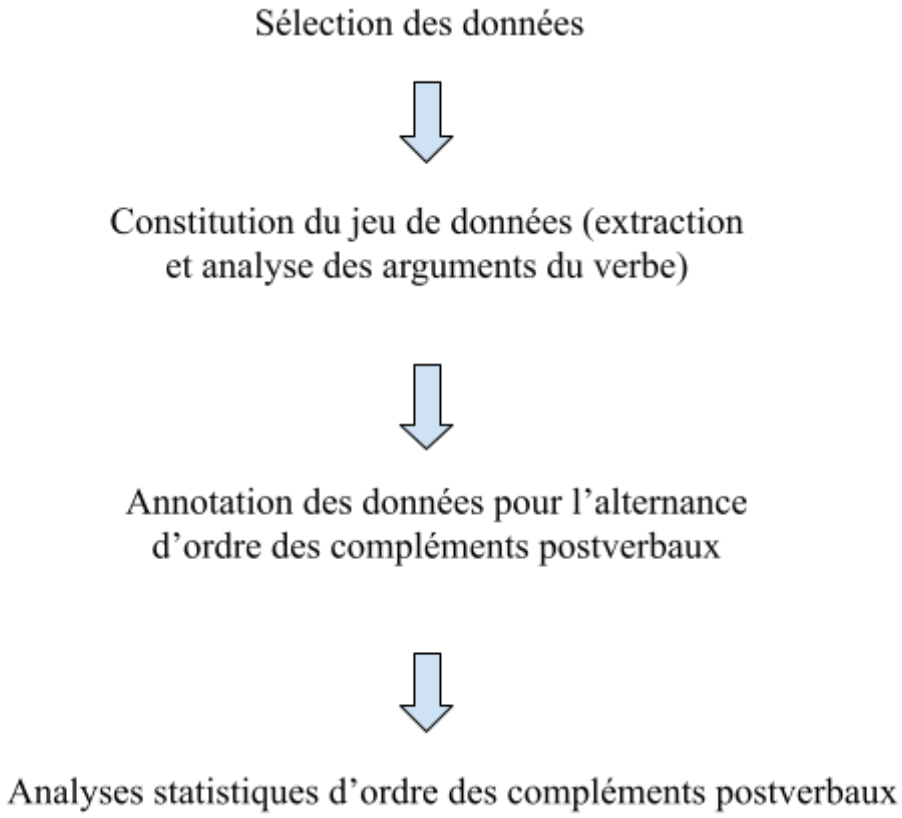


Figure 1 *Les étapes de la démarche expérimentale choisie*

Cette troisième section du mémoire présente quelles sont les données à disposition pour construire le jeu de données sur lequel je me concentrerai dans l'analyse d'ordre des compléments du verbe *donner*. Je commencerai par la présentation du corpus CEFC (section 3.1), puis je présenterai les sous-corpus sélectionnées et les critères d'exclusion dans le cas de certains sous-corpus (section 3.2). Je continuerai par la description des structures-cible pour la construction du jeu de données (section 3.3), suivie par l'illustration de la vue d'ensemble et des limites d'une première exploration des données sélectionnées (section 3.4). A la fin, je mettrai en discussion les différences entre les deux méthodes existantes d'extraction des données : la méthode d'extraction directe depuis la plateforme sur laquelle le corpus se trouve et la méthode d'extraction par l'intermédiaire d'un script informatique (section 3.5).

3.1 Corpus CEFC

Le *Corpus d'Étude pour le Français Contemporain* (*CEFC*) est disponible sur la plateforme *Orféo*[3] (*Outils et Ressources sur le Français Écrit et Oral*) qui a été mise en place grâce à un projet (*ANR 12-CORP-0005*) financé par l'*Agence Nationale de la Recherche* dans le cadre de la campagne Corpus, Données et Outils de la Recherche en Sciences Humaines et Sociales 2011 (Benzitoun et al.,

[3] https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html

2016). Le corpus contient 10 millions de mots, avec une partie orale de 4 millions de mots qui regroupe 14 sous-corpus et une partie écrite de 6 millions de mots qui regroupe 6 sous-corpus. La partie orale comprend des enregistrements récents de locuteurs adultes relevant de situations de parole diverses : conversation, interaction avec des services, prise de parole, réunion, etc. Les discours de 2500 locuteurs différents provenant de l'ensemble des régions de France ainsi que de Suisse et de Belgique ont été recensés dans la base de données *CEFC*. Les sous-corpus oraux disponibles sur la plateforme d'interrogation avec leur taille, leur nombre de mots et leurs situations de communication sont présents dans le tableau suivant[4] :

Corpus	**Taille (mots)**	**Situation de communication**
Corpus de référence du français parlé (CRFP)	374 789	Enregistrements recueillis dans une quarantaine de villes différentes et échantillonnés en fonction de 3 situations de parole et de certaines caractéristiques des locuteurs (niveaux d'études, âge, sexe).
Corpus de français parlé parisien des années 2000 (CFPP)	381 704	Ensemble d'interviews conversationnelles sur les quartiers de Paris et de la proche banlieue.
Corpus de Langue Parlée en Interaction (CLAPI)	210 552	Banque de données multimédia de corpus vidéo et audios enregistrés en situations naturelles dans des contextes variés.
French Oral Narrative (FONC)	131 794	Contes d'une variété de types (fantastiques, merveilleux, facétieux etc.).
VALIBEL	402 285	Enregistrements de productions orales des locuteurs originaires de Bruxelles et de la Wallonie.
C-ORAL-ROM-fr	225 554	Partie française de l'ensemble de corpus comparable de langue spontanée des langues romanes principales français, italien, portugais et espagnol).
FLEURON	30 267	Ressources multimédia représentatives des situations auxquelles les étudiants étrangers seront confrontés lors de leur arrivée dans une université française.

[4] Les informations sont reprises de Bérard (2020).

Traitement de Corpus Oraux en Français (TCOF)	374 789	Enregistrements de corpus d'interactions entre adultes et enfants et des enregistrements d'interactions entre adultes dans différentes situations de communication (conversation, entretien, récit de vie, réunion de travail, etc.).
Réunions de travail	210 552	Enregistrements de réunions de travail.
*Tokyo University of Foreign Studies (*TUFS)	663 742	Partie française d'un corpus d'entretiens multilingues (canadien, espagnol, français, japonais, malaisien, turc).
Corpus de français parlé à Bruxelles (CFPB)	58 688	Semblable au corpus CFPP, mais de Bruxelles.
Corpus Oral de Français de Suisse Romande (OFROM)	258 089	Enregistrements de français parlé en Suisse romande aligné texte/son en situation d'entretien.

Tableau 2 *Les sous-corpus oraux du corpus CEFC*

Selon Benzitoun et al. (2016), l'ensemble des données a été annoté semi-automatiquement en lemmes, catégories grammaticales et fonctions syntaxiques par un analyseur syntaxique. Afin de naviguer dans le corpus, il est possible d'effectuer des recherches simples par concordancier à partir d'une chaîne de caractères (mot ou expression) et des recherches avancées qui peuvent porter sur le lemme, la catégorie grammaticale et les relations de dépendance entre les constituants. La recherche se fait en fonction d'une commande que la plate-forme offre par défaut en fonction de la recherche souhaitée. Les requêtes sont réalisées à l'aide du logiciel *Grew-match* qui utilise des expressions régulières. Un guide qui explique l'utilisation de *Grew-match* est disponible dans la partie *Documentation* de la page.

3.2 Sous-corpus utilisés

Dans ce mémoire, je m'intéresse exclusivement à la partie orale du corpus CEFC qui relève de la langue spontanée parlée par des locuteurs français natifs. Par conséquent, j'ai décidé d'écarter les sous-corpus de la partie écrite, le sous-corpus FLEURON, car les locuteurs sont des étudiants étrangers, le sous-corpus French Oral Narrative, car il ne contient pas de situations qui relèvent de l'oral spontané et le sous-corpus C-ORAL-ROM, car il s'agit de la langue parlée à la radio, ce qui diminue le caractère spontané vu l'organisation du discours dans le cadre d'un tel genre.

Par conséquent, les sous-corpus que j'ai sélectionnés sont :

- *Corpus de référence du français parlé* (CRFP)
- *Corpus de français parlé parisien des années 2000* (CFPP)
- *Corpus de Langue Parlée en Interaction* (CLAPI)
- *VALIBEL*
- *Traitement de Corpus Oraux en Français* (TCOF)
- *Réunions-de-travail*
- *Tokyo University of Foreign Studies (*TUFS)
- *Corpus de français parlé à Bruxelles* (CFPB)

- *Corpus Oral de Français de Suisse Romande* (OFROM)

3.3 Le verbe donner : les réalisations de ses arguments

Cette section présente les réalisations des deux arguments du verbe *donner* présents exclusivement dans la zone postverbale, exclusivement dans la zone préverbale ou dans les deux zones.

Puisque le verbe *donner* permet d'analyser le phénomène de l'alternance des compléments et offre la possibilité d'observer une multitude des réalisations de ses arguments qui de fois impliquent le facteur pronominalité, j'ai choisi de me concentrer dans ce mémoire sur ce verbe qui implique la préposition *à* à l'intérieur de l'un de ses arguments. *Donner* est un verbe assez fréquent (Il se retrouve dans 3003 phrases parmi les 371.927 phrases du corpus CEFC sélectionné, soit 0,81% des phrases du corpus entier.) et la probabilité de rencontrer toutes ces structures est plus élevée que dans les cas des autres verbes ditransitifs moins fréquents, comme *offrir* (Il se retrouve dans 184 phrases, soit 0.05% des phrases du corpus entier). Par conséquent, le verbe *donner* offre une bonne image de la multitude des réalisations de ses arguments dans le français oral spontané.

Selon la zone occupée par rapport au verbe, je distingue trois grands types de structures-cibles qui se retrouvent dans le jeu de données constitue. Le premier type (A) suppose la réalisation des arguments du verbe dans une zone strictement postverbale et l'un des arguments est un syntagme nominal à tête nominale ou à tête pronominale que je noterai *SN* et respectivement *SN(pron)* et l'autre argument est un syntagme prépositionnel (dont la tête est la préposition *à*) que je noterai *SP*. Le deuxième grand type (B) suppose la réalisation des arguments du verbe autant dans la zone préverbale par l'intermédiaire d'un pronom clitique et dans la zone postverbale un syntagme prépositionnel ou un syntagme nominal à tête nominale ou pronominale suivi ou non d'un syntagme prépositionnel disloqué. Les deux derniers sous-types supposent la réalisation des deux arguments du verbe dans la zone préverbale et la présence d'un des arguments aussi dans la zone postverbale sous forme d'un syntagme nominal ou prépositionnel. Le troisième grand type (C) suppose la réalisation des arguments du verbe strictement dans une zone préverbale par l'intermédiaire des deux pronoms clitiques. Chaque type et sous-type sont présentés et accompagnés par un exemple dans les lignes qui suivent :

A. Le verbe *donner* qui sous-catégorise ses deux arguments strictement dans la zone postverbale

Les deux arguments qui sont sous-catégorisés par le verbe *donner* dans la zone postverbale peuvent générer 2 types des structures indépendamment de l'ordre des arguments :

a) donner + SN + SP

   ex. Après il euh il fallait bon donner [des cours]$_{SN}$ [à des élèves qui venaient pas tout le temps]$_{SP}$.

b) donner + SN(pron) + SP

   ex. Voilà il a donné [ça]$_{SN(pron)}$ [à toute la classe]$_{SP}$.

B. Le verbe *donner* qui sous-catégorise ses deux arguments dans les zones préverbale et postverbale

Les deux arguments qui sont sous-catégorisés par le verbe *donner* dans les zones préverbale et postverbale peuvent générer indépendamment de l'ordre des arguments les types suivants de structures :

a) CLI + donner + SN (2 arguments du verbe, l'un réalisé sous forme du clitique et l'autre dans la zone postverbale)

ex. Écoute-moi je vais [te]$_{CLI}$ donner [un exemple]$_{SN}$ aussi.

b) CLI + donner + SN(pron) (2 arguments du verbe, l'un réalisé sous forme du clitique et l'autre dans la zone postverbale)

ex. Ah elle [t']$_{CLI}$ a donné [ça]$_{SN(pron)}$.

c) CLI + donner + SP (2 arguments du verbe, l'un réalisé sous forme du clitique et l'autre dans la zone postverbale)

ex. Oui ben enfin il aura il il va y avoir de la jalousie parce qu' on pourra pas [les] donner [à tout le monde] euh.

d) CLI + donner + SN + SP(disloqué) (2 arguments du verbe, l'un réalisé sous forme du clitique qui redouble l'un des arguments du verbe et un autre dans la zone postverbale)

ex. Il [te]$_{CLI}$ donne [des adresses]$_{SN}$ [à toi]$_{SP}$.

e) CLI + donner + SN(pron) + SP(disloqué) (2 arguments du verbe, l'un réalisé sous forme du clitique et sous forme de redoublement et un autre dans la zone postverbale)

ex. Je [leur]$_{CLI}$ ai donné [tout]$_{SN(pron)}$ [à eux]$_{[SP]}$.

f) CLI1 + CLI2 + donner + SN(disloqué) (2 arguments du verbe réalisés dans la zone préverbale sous forme des clitiques dont l'un est aussi présent dans la zone postverbale sous forme de SN)

ex. Bah je vais [vous]$_{CLI1}$ [le]$_{CLI2}$ donner [le programme]$_{SN}$.

g) CLI1 + CLI2 + donner + SN(pron, disloqué) (2 arguments du verbe réalisés dans la zone préverbale sous forme des clitiques dont l'un est aussi présent dans la zone postverbale sous forme de SN à tête pronominale)

ex. Moi je veux bien [vous]$_{CLI1}$ [le]$_{CLI2}$ donner [le mien]$_{SN(pron)}$.

h) CLI1 + CLI2 + donner + SP(disloqué) (2 arguments du verbe réalisés dans la zone préverbale sous forme des clitiques dont l'un est aussi présent dans la zone postverbale sous forme de SP)

ex. Je vais [te]$_{CLI1}$ [le]$_{CLI2}$ donner [à toi]$_{SP}$.

C) Le verbe *donner* qui sous-catégorise ses deux arguments strictement dans la zone préverbale

i) CLI1 + CLI2 + donner (Les deux arguments du verbe sont réalisés sous forme de pronoms clitiques.)

ex. Après il [me]$_{CLI1}$ [les]$_{CLI2}$ donne.

Le verbe de transfert *donner* (se construit avec la préposition *à*) est un verbe fréquent et la probabilité de rencontrer toutes ces constructions qui supposent la réalisation de ses arguments avant et/ou après le verbe assez élevée.

La section suivante décrit les procédures qui ont été mises en place afin de construire le jeu de données en se basant sur ces structures-cible.

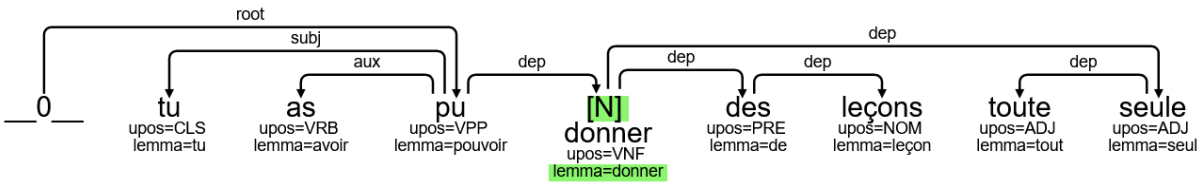
3.4 Une première exploration des données

Les données du corpus CEFC sont disponibles sur la plateforme d'interrogation *Orféo* et elles peuvent être récupérées en formulant des requêtes à l'aide des expressions régulières.

Afin d'extraire les données dont j'ai besoin pour mes analyses, j'ai utilisé des modèles de requêtes combinées déjà proposées par le *Grew-match* et qui intègrent des expressions régulières.

Il existe la possibilité de faire des requêtes plus génériques et des requêtes plus spécifiques. Les requêtes génériques, comme la recherche de toutes les phrases qui contiennent le lemme *donner* (pattern { N [lemma="donner"] }), apportent du bruit, car parmi toutes les phrases contenant le verbe cible, il existe des phrases qui ne font pas l'objet du phénomène syntaxique ciblé dans lequel les verbes *donner* a deux arguments, comme dans l'exemple suivant :

17)

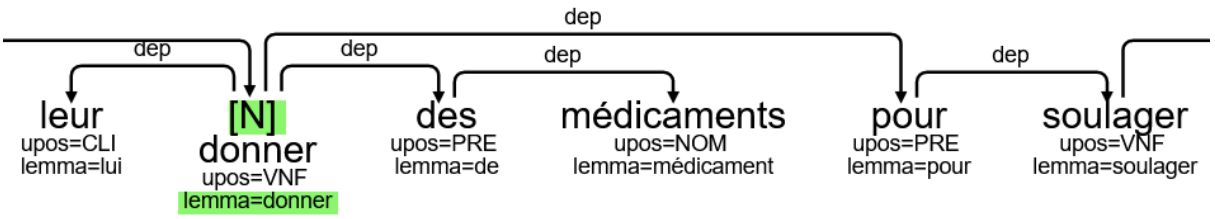


(Exemple tiré du corpus *OFROM*)

Dans cet exemple, le verbe donner n'a qu'un argument, plus exactement le SN *des leçons*.

Le sous-corpus *OFROM* contient 257 phrases avec le lemme *donner*. Parmi ces phrases, il y a des phrases qui incluent le phénomène syntaxique étudié et contiennent l'un des structures-cible, comme :

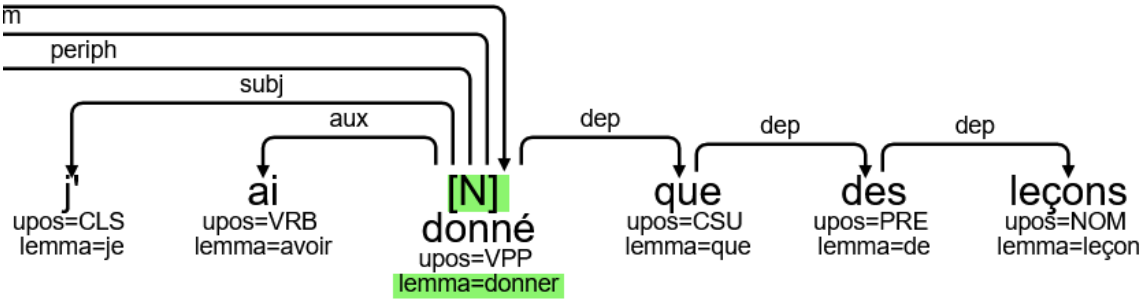
18)



Ainsi, l'un des arguments du verbe est présent dans la zone préverbale sous forme du pronom clitique *leur* et l'autre argument se retrouve dans la zone postverbale sous forme du nom *médicaments*.

Cependant, parmi ces 257 phrases, il existe aussi des phrases qui ne correspondent pas au phénomène syntaxique étudié, comme :

19)

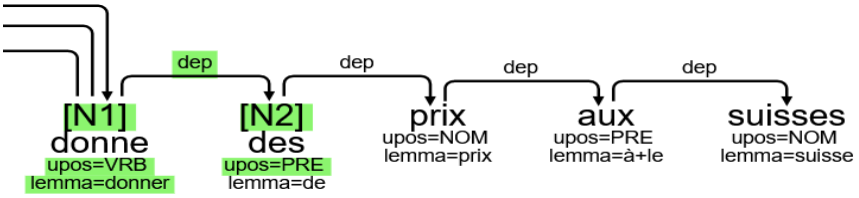


(Exemple trouvé dans le sous-corpus *OFROM*)

Ainsi, il s'avère difficile de classer les types de phrase par l'intermédiaire des requêtes sur la plateforme, car il est nécessaire de développer des conditions complexes pour distinguer tous ces phénomènes.

Un autre problème qui apparaît est représenté par les phrases qui sont mal annotées de manière récurrente, comme dans le cas de l'exemple suivant :

20)



(Exemple trouvé dans le sous-corpus *OFROM*)

C'est ainsi qu'il existe des cas où le déterminant *des* est annoté comme préposition et le syntagme prépositionnel *aux suisses* est rattaché à la tête du syntagme nominal *des prix* au lieu d'être rattaché au verbe.

Pour avoir une vision plus claire de la proportion des phrases pertinentes dans chaque sous-corpus sélectionné, j'ai effectué une requête spécifique qui suit les règles standard d'annotation et qui est valable pour les cas où les deux arguments du verbe *donner* se retrouvent exclusivement dans la zone postverbale. Il s'agit plus précisément des situations où le verbe gouverne la tête du syntagme nominal et la tête du syntagme prépositionnel (*pattern {N1 [lemma ="donner", upos= re"V.*"]; N2 [upos=NOM|PRO]; N3[upos="PRE", lemma="à"] ; N1 -[dep]-> N2 ; N1 -[dep]-> N3 }*). Puis, j'ai compté manuellement les phrases en fonction de leur pertinence en ce qui concerne l'annotation. Par conséquent, les résultats de l'extraction sont visibles dans la Figure 2.

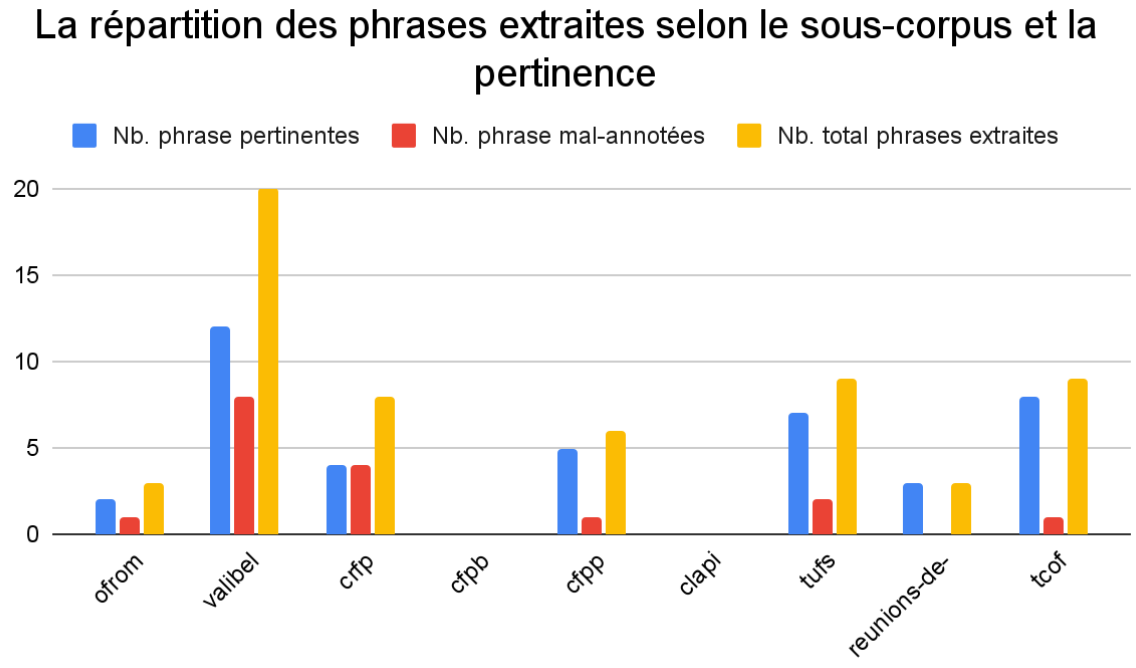


Figure 2 *La répartition des phrases extraites selon le sous-corpus et la pertinence de l'annotation*

Ainsi, parmi les phrases qui sont pertinentes pour le phénomène syntaxique étudié et suivent la requête indiquée précédemment, il existe du bruit. La situation en pourcentage des phrases qui n'ont pas été annotées correctement au niveau de chaque sous-corpus sélectionné se présente de la manière suivante :

- *Corpus de référence du français parlé* (CRFP) : 50%
- *Corpus de français parlé parisien des années 2000* (CFPP) : 16.67%
- *Corpus de Langue Parlée en Interaction* (CLAPI) : 0%
- *VALIBEL* : 40%
- *Traitement de Corpus Oraux en Français* (TCOF) : 11.11%
- *Réunions-de-travail* : 0%
- *Tokyo University of Foreign Studies (*TUFS) : 22.22%
- *Corpus de français parlé à Bruxelles* (CFPB) : 0%
- *Corpus Oral de Français de Suisse Romande* (OFROM) : 33.33 %

Vu ces résultats, le bruit s'avère assez important et le fait que la requête n'a trouvé aucune phrase dans les sous-corpus *CFPB* et *CLAPI* peut être un indice du silence existant. Il est difficile d'évaluer le silence, car au-delà des erreurs qui sont récurrentes au niveau de l'annotation, il peut exister des erreurs isolées qui poseront des problèmes à l'extraction des données.

Pour résumer, il existe de nombreux problèmes d'annotation qui génèrent du bruit et peuvent générer du silence et il est impossible d'estimer leur proportion par l'intermédiaire des extractions depuis la plateforme. Il sera possible de mieux évaluer les deux mesures à l'aide d'un script Python qui représente une méthode plus efficace (sous-section 3.5) en choisissant un échantillon à partir duquel un Gold standard est construit (sous-section 4.3).

3.5 Plateforme vs. Script

Afin de construire le jeu de données nécessaire pour l'analyse du phénomène syntaxique d'alternance d'ordre des compléments du verbe, il existe deux méthodes d'extraction des phrases, plus exactement

par l'intermédiaire de la plateforme d'interrogation sur laquelle le corpus est disponible et par l'intermédiaire d'un extracteur réalisé dans le cadre d'un script informatique.

Contrairement à l'utilisation de la première méthode, l'utilisation d'un extracteur permet une extraction plus efficace au niveau du temps, car il est possible d'obtenir des résultats de tous les sous-corpus à la fois, ce qui est impossible dans le cas de la plateforme d'interrogation (L'interrogation se fait sous-corpus par sous-corpus.).

Un autre avantage de la deuxième méthode est le fait de pouvoir extraire aussi les phrases qui font l'objet d'une annotation incorrecte en observant les tendances générales des erreurs et en adaptant l'extracteur.

Cependant, l'extracteur construit ne permettra pas malheureusement d'éviter intégralement les problèmes d'annotation observés dans la sous-section antérieure, car il existe des erreurs d'annotation qui ne font pas l'objet d'une tendance générale.

La section suivante présente la démarche choisie pour l'extraction des données utilisées dans l'analyse des arguments du verbe *donner*.

4 Constitution du jeu de données

Cette section est consacrée à la présentation du jeu des données obtenu après plusieurs manipulations réalisées à partir du corpus CEFC. Elle commence par l'illustration des transformations et procédures qui ont été utilisées à partir du corpus CEFC afin d'arriver à la constitution du jeu de données (sous-section 4.1) et la méthode d'évaluation de l'extracteur de données qui a été mise en place par rapport à un Gold standard que j'ai réalisé manuellement (sous-section 4.2). Dans un dernier temps, elle présente le jeu de données créé pour le corpus entier sélectionné (sous-section 4.3) et les limites dépassées et restantes à la suite de l'emploi d'une telle méthode de constitution du jeu de données qui va faire l'objet d'une procédure d'annotation et des analyses statistiques (sous-section 4.4).

4.1 Extraction des données : sous-corpus *CRFP*

Cette sous-section présente d'abord les étapes de la procédure qui a été nécessaire afin de construire un jeu de données final fiable et dans un deuxième temps les premières extractions de données à partir d'un échantillon choisi du corpus sélectionné, plus exactement les phrases contenant le verbe *donner* du sous-corpus *CRFP*.

Chaîne de traitement

Puisque la méthode du script s'est avérée plus efficace au niveau du temps d'extraction des données, mais les annotations imposent des limites insurmontables afin d'assurer un contrôle minimal sur la qualité de l'extraction (sous-section 3.4), je me suis orientée vers la une re-annotation des données à l'aide du parseur *Stanza*[5] (Qi et al., 2020) qui est une version plus moderne du parser *MACAON*[6] (Nasr et al., 2011).

Cette ré-analyse du corpus implique le passage du format initial dans lequel le corpus a été disponible (orfeo) au format généré par le parser *Stanza* (CoNNL-U). Le format offert par *Stanza* est un format similaire au format orfeo, mais il contient des informations morpho-syntaxiques plus exactes et plus avancées par rapport au dernier format mentionné.

Stanza est un parser qui assure une très bonne qualité au niveau de l'annotation vu son rendement par rapport à d'autres parser, comme UDPipe et spaCY à la suite des tests effectués sur les banques d'arbres disponibles qui suivent les normes de Universal Dependencies (UD)[7] (Qi et. al, 2020). Par conséquent, les relations de dépendance devraient se manifester de la manière suivante :

- La tête du SN du verbe doit dépendre directement du verbe

21)

```
# text = alors du coup moi j' ai essayé de donner son sac de congrès à un gars
# sent_id = cefc-ofrom-unine08a03m-128
1       alors   alors   ADV     _       _       8       advmod  _       start_char=12147|end_char=12152|ner=O
2-3     du      _       _       _       _       _       _       _       start_char=12153|end_char=12155|ner=O
2       de      de      ADP     _       _       4       case    _       _
3       le      le      DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       2       fixed   _       _
4       coup    coup    NOUN    _       Gender=Masc|Number=Sing 1       fixed   _       start_char=12156|end_char=12160|ner=O
5       moi     lui     PRON    _       Number=Sing|Person=1|PronType=Prs       4       nmod    _       start_char=12161|end_char=12164|ner=O
6       j'      il      PRON    _       Number=Sing|Person=1|PronType=Prs       8       nsubj   _       start_char=12165|end_char=12167|ner=O
7       ai      avoir   AUX     _       Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin   8       aux:tense       _       start_char=12168|end_char=12170|ner=O
8       essayé  essayer VERB    _       Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part        0       root    _       start_char=12171|end_char=12177|ner=O
9       de      de      ADP     _       _       10      mark    _       start_char=12178|end_char=12180|ner=O
10      donner  donner  VERB    _       VerbForm=Inf    8       xcomp   _       start_char=12181|end_char=12187|ner=O
11      son     son     DET     _       Number=Sing|Number[psor]=Sing|Person[psor]=3|Poss=Yes|PronType=Prs       12      det     _       start_char=12188|end_char=12191|ner=O
12      sac     sac     NOUN    _       Gender=Masc|Number=Sing 10      obj     _       start_char=12192|end_char=12195|ner=O
13      de      de      ADP     _       _       14      case    _       start_char=12196|end_char=12198|ner=O
14      congrès congrès NOUN    _       Gender=Masc|Number=Sing 12      nmod    _       start_char=12199|end_char=12206|ner=O
15      à       à       ADP     _       _       17      case    _       start_char=12207|end_char=12208|ner=O
16      un      un      DET     _       Definite=Ind|Gender=Masc|Number=Sing|PronType=Art       17      det     _       start_char=12209|end_char=12211|ner=O
17      gars    gars    NOUN    _       Gender=Masc|Number=Sing 10      obl:arg _       start_char=12212|end_char=12216|ner=O
```

[5] https://stanfordnlp.github.io/stanza/

[6] Ce parser est le parser utilisé pour l'étiquetage morpho-syntaxique du corpus *CEFC*.

[7] https://universaldependencies.org/
UD est un projet de recherche collaboratif qui vise à développer des normes universelles pour l'annotation de la syntaxe des langues naturelles.

Dans cet exemple, la tête du SN *sac* dépend directement du verbe *donner*.

- Le nom qui est rattaché à la tête du SP (la préposition) doit dépendre directement du verbe

22)

```
# text = après il euh il fallait bon donner des cours à des élèves qui venaient pas tout le temps
# sent_id = cefc-ofrom-unine08a10m-123
1	après	après	ADP	_	_	3	mark	_	start_char=31875|end_char=31880|ner=O
2	il	il	PRON	_	Gender=Masc|Number=Sing|Person=3|PronType=Prs	3	expl:subj	_	start_char=31881|end_char=31883|ner=O
3	euh	avoir	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	5	advcl	_	start_char=31884|end_char=31887|ner=O
4	il	il	PRON	_	Gender=Masc|Number=Sing|Person=3|PronType=Prs	5	expl:subj	_	start_char=31888|end_char=31890|ner=O
5	fallait	falloir	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin	0	root	_	start_char=31891|end_char=31898|ner=O
6	bon	bon	ADJ	_	Gender=Masc|Number=Sing	5	nsubj	_	start_char=31899|end_char=31902|ner=O
7	donner	donner	VERB	_	VerbForm=Inf	5	ccomp	_	start_char=31903|end_char=31909|ner=O
8	des	un	DET	_	Definite=Ind|Number=Plur|PronType=Art	9	det	_	start_char=31910|end_char=31913|ner=O
9	cours	cours	NOUN	_	Gender=Masc|Number=Plur	7	obj	_	start_char=31914|end_char=31919|ner=O
10	à	à	ADP	_	_	12	case	_	start_char=31920|end_char=31921|ner=O
11	des	un	DET	_	Definite=Ind|Number=Plur|PronType=Art	12	det	_	start_char=31922|end_char=31925|ner=O
12	élèves	élève	NOUN	_	Gender=Masc|Number=Plur	7	obl:arg	_	start_char=31926|end_char=31932|ner=O
13	qui	qui	PRON	_	PronType=Rel	14	nsubj	_	start_char=31933|end_char=31936|ner=O
14	venaient	venir	VERB	_	Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin	12	acl:relcl	_	start_char=31937|end_char=31945|ner=O
15	pas	pas	ADV	_	Polarity=Neg	14	advmod	_	start_char=31946|end_char=31949|ner=O
16	tout	tout	ADJ	_	Gender=Masc|Number=Sing	18	amod	_	start_char=31950|end_char=31954|ner=O
17	le	le	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	18	det	_	start_char=31955|end_char=31957|ner=O
18	temps	temps	NOUN	_	Gender=Masc|Number=Sing	14	obj	_	start_char=31958|end_char=31963|ner=O
```

Dans cet exemple, le nom *élèves* dépend directement du verbe *donner*.

- Le pronom clitique doit dépendre directement du verbe

23)

```
# text = et ça je lui donne le détail
# sent_id = cefc-cfpp-Bernard_Rosier_H_60_Micheline_Rosier_58_12e-2561
1	et	et	CCONJ	_	_	5	cc	_	start_char=267774|end_char=267776|ner=O
2	ça	ça	PRON	_	Gender=Masc|Number=Sing|Person=3|PronType=Dem	5	nsubj	_	start_char=267777|end_char=267779|ner=O
3	je	il	PRON	_	Number=Sing|Person=1|PronType=Prs	5	nsubj	_	start_char=267780|end_char=267782|ner=O
4	lui	lui	PRON	_	Gender=Masc|Number=Sing|Person=3|PronType=Prs	5	iobj	_	start_char=267783|end_char=267786|ner=O
5	donne	donner	VERB	_	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin	0	root	_	start_char=267787|end_char=267792|ner=O
6	le	le	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	7	det	_	start_char=267793|end_char=267795|ner=O
7	détail	détail	NOUN	_	Gender=Masc|Number=Sing	5	obj	_	start_char=267796|end_char=267802|ner=O
```

Dans ce cas, le pronom clitique *lui* est directement rattaché au verbe.

Afin d'assurer plus de contrôle sur le fonctionnement de la chaîne de traitement que j'ai réalisée, j'ai commencé à appliquer la méthodologie finale sur un échantillon de la partie orale du corpus CEFC, notamment le sous-corpus *CRFP*. Mon choix s'est porté sur ce corpus, car le verbe *donner* se retrouve dans 349 phrases parmi les 41533 phrases du sous-corpus entier, soit 0.84% du nombre total des phrases et 11.62% du nombre total des phrases contenant le verbe *donner*, ce qui garantit une diversité des réalisations et des phénomènes qui peuvent apparaître à l'intérieur des phrases qui m'intéressent. La chaîne de traitement que j'ai réalisée comporte trois étapes :

a) Extraction des phrases contenant le lemme *donner*

Afin de rendre plus efficace le traitement de *Stanza*, j'ai extrait toutes les phrases contenant le lemme *donner*. Puisque le parseur initial a bien étiqueté et lemmatisé les verbes, j'ai pu récupérer, à l'aide d'un script, toutes les phrases concernées dans un format txt.

b) Re-annotation des données par l'intermédiaire de l'analyse réalisée par *Stanza*

J'ai pu obtenir le corpus des phrases contenant le lemme *donner* sous le format CONNL-U grâce à l'analyse du texte brut par le parseur *Stanza*.

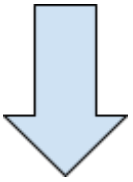
c) Adaptation du script d'extraction utilisé (sous-section 3.5)

Cette étape a supposé l'adaptation des éléments du script déjà utilisés au niveau des étiquettes morpho-syntaxiques et aux relations de dépendance qui sont utilisées par *Stanza*, notamment selon les normes de UD. Le but de cette étape a été d'extraire les structures-cible pour mon étude grâce à des données qui ne comportent pas de limites insurmontables au niveau de l'annotation.

Pour résumer, les étapes de la chaîne du traitement que j'ai réalisée sont visibles dans le schéma suivant qui illustre tous les états d'une phrase présente dans le sous-corpus *CRFP* sélectionné :

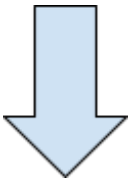
Format orfeo

```
# sent_id = cefc-crfp-PUB-BES-1-304
# text = ça aura donné une bonne raison à quelqu' un de le couper cet arbre-là par contre
1       ça      ça      PRO     PRO     _       3       subj    _       _       1369.939941     1370.170044     L1
2       aura    avoir   VRB     VRB     _       3       aux     _       _       1370.180054     1370.469971     L1
3       donné   donner  VPP     VPP     _       0       root    _       _       1370.479980     1370.680054     L1
4       une     un      DET     DET     _       6       spe     _       _       1370.689941     1370.849976     L1
5       bonne   bon     ADJ     ADJ     _       6       dep     _       _       1370.859985     1371.010010     L1
6       raison  raison  NOM     NOM     _       3       dep     _       _       1371.020020     1371.359985     L1
7       à       à       PRE     PRE     _       6       dep     _       _       1371.369995     1371.390015     L1
8       quelqu' un      quelqu'unf      PRO     PRO     _       7       dep     _       _       1371.459961     1371.729980     L1
9       de      de      PRE     PRE     _       6       dep     _       _       1371.900024     1371.920044     L1
10      le      le      CLI     CLI     _       11      dep     _       _       1371.930054     1371.949951     L1
11      couper  couper  VNF     VNF     _       9       dep     _       _       1371.959961     1372.250000     L1
12      cet     ce      DET     DET     _       13      spe     _       _       1372.260010     1372.420044     L1
13      arbre-là        arbre-là        NOM     NOM     _       11      dep     _       _       1372.430054     1372.719971     L1
14      par     par     PRE     PRE     _       11      dep     _       _       1373.119995     1373.270020     L1
15      contre  contre  PRE     PRE     _       14      para    _       _       1373.280029     1373.479980     L1
```



Format txt

```
ça aura donné une bonne raison à quelqu' un de le couper cet arbre -là par contre
```



Format CoNNL-U

```
# text = ça aura donné une bonne raison à quelqu' un de le couper cet arbre-là par contre
# sent_id = cefc-crfp-PUB-BES-1-304
1       ça      ça      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem    3       nsubj   _       start_char=77394|end_char=77396|ner=0
2       aura    avoir   AUX     _       Mood=Ind|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin    3       aux:tense       _       start_char=77397|end_char=77401|ner=0
3       donné   donner  VERB    _       Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part        0       root    _       start_char=77402|end_char=77407|ner=0
4       une     un      DET     _       Definite=Ind|Gender=Fem|Number=Sing|PronType=Art        6       det     _       start_char=77408|end_char=77411|ner=0
5       bonne   bon     ADJ     _       Gender=Fem|Number=Sing  6       amod    _       start_char=77412|end_char=77417|ner=0
6       raison  raison  NOUN    _       Gender=Fem|Number=Sing  3       obj     _       start_char=77418|end_char=77424|ner=0
7       à       à       ADP     _       _       9       case    _       start_char=77425|end_char=77426|ner=0
8       quelqu' quelqu't        ADJ     _       Gender=Masc|Number=Sing 9       amod    _       start_char=77427|end_char=77434|ner=0
9       un      un      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Ind    3       obl:arg _       start_char=77435|end_char=77437|ner=0
10      de      de      ADP     _       _       12      mark    _       start_char=77438|end_char=77440|ner=0
11      le      le      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Prs    12      obj     _       start_char=77441|end_char=77443|ner=0
12      couper  couper  VERB    _       VerbForm=Inf    9       acl     _       start_char=77444|end_char=77450|ner=0
13      cet     ce      DET     _       Gender=Masc|Number=Sing|PronType=Dem    14      det     _       start_char=77451|end_char=77454|ner=0
14      arbre   arbre   NOUN    _       Gender=Masc|Number=Sing 12      obj     _       start_char=77455|end_char=77460|ner=0
15      -là     là      ADV     _       _       14      advmod  _       start_char=77460|end_char=77463|ner=0
16      par     par     ADP     _       _       12      advmod  _       start_char=77464|end_char=77467|ner=0
17      contre  contre  NOUN    _       _       16      fixed   _       start_char=77468|end_char=77474|ner=0
```

Figure 3 *Chaîne de traitement pour une phrase du sous-corpus CRFP*

C’est ainsi que le texte de base a pu être ré-analysé et cela a permis d'écarter de nombreuses erreurs d’annotation initiale que je détaillerai dans la sous-section 4.5.

4.2 Les premières extractions

Une fois mise en place la chaîne de traitement élaborée et après avoir amélioré le script de l'extraction au fur et à mesure en regardant la sortie pour les phrases dont les arguments du verbe *donner* se réalisent exclusivement dans la zone postverbale (type A), j'ai obtenu les résultats suivants :

Type de structure	Nombre de phrases	% du nombre de phrases contenant le verbe *donner* du corpus *CRFP*	% du nombre de phrases contenant le verbe *donner* du corpus *CRFP* [+type A]
donner + SN + SP	22	6.5%	100%
donner + SN(pron) + SP	0	0%	0%

Tableau 3 *Les résultats chiffrés des données de type A le sous-corpus CFRP*

C'est ainsi que les phrases où les arguments du verbe *donner* sont réalisés exclusivement dans la zone postverbale représentent 6.5% des phrases du corpus *CRFP* sélectionné (phrase contenant le verbe *donner*). À l'intérieur du grand type A, il existe que des phrases qui contiennent la structure donner + SN + SP.

Pour ce qui est des phrases dont les arguments du verbe *donner* se réalisent aussi bien dans la zone préverbale que dans la zone postverbale (type B), les résultats sont donnés dans le *Tableau 4* :

Type de structure	Nombre de phrases	% du nombre de phrases contenant le verbe donner du corpus *CRFP*	% du nombre de phrases contenant le verbe donner du corpus *CRFP* [+ type B]
CLI + donner + SN	110	32.54%	94.83%
CLI + donner + SN(pron)	0	0%	0%
CLI + donner + SP	3	0.89%	2.59%
CLI + donner + SN + SP(disloqué)	3	0.89%	2.59%
CLI + donner + SN(pron) + SP(disloqué)	0	0%	0%

CLI1 + CLI2 + donner + SN(disloqué)	0	0%	0%
CLI1 + CLI2 + donner + SN(pron, disloqué)	0	0%	0%
CLI1 + CLI2 + donner + SP(disloqué)	0	0%	0%

Tableau 4 *Les résultats chiffrés des données de type B le sous-corpus CFRP*

On observe que la structure de type CLI + donner + SN est la plus présente parmi les phrases dans lesquelles la réalisation des arguments du verbe se fait dans les deux zones (94.83%). Cependant, la plupart des structures ne se retrouvent pas dans le sous-corpus *CRFP*. Le tableau montre également qu'il existe 116 phrases de type B, soit 34.32% du corpus *CRFP* sélectionné.

Quant aux phrases où les arguments du verbe *donner* se réalisent exclusivement dans la zone préverbale (type C), les résultats sont donnés dans le *Tableau 5* :

Type de structure	Nombre de phrases	% du nombre de phrases contenant le verbe donner du corpus *CRFP*	% du nombre de phrases contenant le verbe donner du corpus *CRFP* [+type C]
CLI1 + CLI2 + donner	1	0.30%	100%

Tableau 5 *Les résultats chiffrés des données de type C le sous-corpus CFRP*

A la suite de l'extraction, il existe une seule phrase où les arguments du verbe *donner* sont présents exclusivement dans la zone préverbale sous forme des pronoms politiques.

La répartition selon le grand type de structures dans l'échantillon sélectionné sélectionne (constitué par les phrases contenant le lemme donner du sous-corpus *CRFP*) est visible dans le *Tableau 6* :

Grand type de structure	% du nombre des phrases contenant le verbe donner dans le sous-corpus *CRFP*
Type A	6.5%
Type B	34.32%
Type C	0.30%
Autres phrases	59.17%

Tableau 6 *La répartition des données selon les grands types de structure dans l'échantillon*

Les phrases qui sont pertinentes dans l'analyse de l'alternance d'ordre des compléments postverbaux en français oral spontané constituent 6.5% parmi toutes les phrase de l'échantillon sélectionné et la plus grande proportion reste composée par les autres phrases qui ne font pas l'objet de ce phénomène syntaxique (59.17%). La structure de type B qui présente la réalisation des arguments du verbe *donner* dans les deux zones domine les autres deux types de structure. Au sens large, les structures appartenant au grand type B et au grand type C (prises ensemble) sont plus fréquentes que ceux qui ne contiennent pas de pronoms clitiques.

La sous-section suivante présente la méthode d'évaluation de l'extracteur mis en place et le score obtenu de différentes mesures spécifiques au domaine du TAL. Une fois que les résultats de l'évaluation indiquent de bons scores, l'extracteur pourra être utilisé pour constituer le jeu de données utilisé dans mes analyses sur l'alternance des compléments du verbe *donner*.

4.3 Évaluation de l'extraction : la méthode du Gold standard

La méthode de l'élaboration d'un Gold standard est une méthode fréquemment utilisée pour évaluer le rendement d'un système dans le domaine du Traitement Automatique des Langues (TAL) et suppose un travail manuel (Adam et al., 2013). Bien que cette méthode ne puisse pas fournir la qualité de l'extraction dans sa globalité une fois l'extraction appliquée au niveau de tout le corpus sélectionné, elle reste un bon indicateur dans l'amélioration de la démarche mise en place.

Afin d'assurer un contrôle du rendement de la chaîne de traitement et de la qualité de l'extraction des données, j'ai élaboré manuellement un gold standard à partir d'un échantillon du corpus total sélectionné, plus exactement les phrases contenant le verbe *donner* du sous-corpus CFRP. Vu sa taille, il représente un bon outil de contrôle pour observer les réalisations des arguments du verbe *donner* et les situations qui peuvent apparaître dans ce contexte. Cette démarche est également nécessaire afin de réussir une annotation aussi bonne que possible au niveau du jeu de données final.

C'est ainsi que le Gold contient 21 phrases dans lesquelles les arguments du verbe *donner* sont réalisés strictement dans la zone postverbale, 118 phrases dans lesquelles les arguments du verbe *donner* sont réalisés autant dans la zone préverbale que dans la zone postverbale et 1 phrase dans laquelle les arguments du verbe *donner* sont réalisés strictement dans la zone préverbale.

Après avoir amélioré l'extracteur mis en place en regardant constamment la sortie, j'ai calculé les mesures qui s'appliquent d'habitude à une telle méthode d'évaluation, plus exactement la précision, le rappel et la F-mesure (ou la mesure F1). La précision représente le nombre des phrases pertinentes (qui se retrouvent dans le gold standard) parmi les phrases que le système a donné en sortie, tandis que la précision représente le nombre de phrases pertinentes parmi la totalité des phrases que le système aurait dû donner en sortie. La F-mesure peut être calculée après avoir calculé la précision et le rappel, car elle représente la moyenne harmonique entre les deux mesures.

Comme il n'existe pas d'occurrences de type donner + SN(pron) + SP, les scores obtenus d'une manière générale pour les deux grand types sont les suivants :

Type de structure	Précision	Rappel	F-mesure
A	88%	100%	0.93
B	98%	98%	0.98
C	100%	100%	1

Tableau 7 *Résultats de mesures dans l'évaluation en fonction de grands types*

C'est ainsi que dans le cas de la structure qui contient les arguments du verbe *donner* strictement dans la zone postverbale, l'extracteur a pu identifier toutes les phrases, mais la précision n'a pas reçu un score si haut vu la présence de certaines erreurs d'annotation qui ont engendré du bruit une fois les bonnes phrases identifiées. Puisque la F-mesure est une mesure qui pénalise les systèmes qui privilégient un score plutôt qu'un autre, tout comme le cas de mon approche (J'ai privilégié le rappel.), le score obtenu est un peu plus modeste, mais reste un bon score. Cependant, pour ce qui est des phrases de type B et C, les scores des 3 mesures sont très hauts.

Bien que l'analyse morpho-syntaxique de Stanza soit meilleure que celle réalisée par le parser *MACAON* (Nasr et al., 2011), il reste toujours des erreurs d'annotation au niveau des relations de dépendance qui peuvent constituer une source du silence. Parmi les erreurs récurrentes d'annotation, on relève :

- le nom dépendant de la préposition *à* est rattaché à la tête du syntagme nominal au lieu d'être rattaché au verbe, comme dans l'exemple suivant :

24)

```
# text = en fait ça en fait ça donne confiance aux gens
# sent_id = cefc-crfp-PRI-PNO-2-433
1       en      en      ADP     _       _       5       advmod  _       start_char=85112|end_char=85114|ner=O
2       fait    fait    NOUN    _       Gender=Masc|Number=Sing 1       fixed   _       start_char=85115|end_char=85119|ner=O
3       ça      ça      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem   5       nsubj   _       start_char=85120|end_char=85122|ner=O
4       en      en      PRON    _       Person=3|PronType=Prs   5       iobj    _       start_char=85123|end_char=85125|ner=O
5       fait    faire   VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root    _       start_char=85126|end_char=85130|ner=O
6       ça      ça      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem   7       nsubj   _       start_char=85131|end_char=85133|ner=O
7       donne   donner  VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   5       conj    _       start_char=85134|end_char=85139|ner=O
8       confiance       confiance       NOUN    _       Gender=Fem|Number=Sing  7       obj:lvc _       start_char=85140|end_char=85149|ner=O
9-10    aux     _       _       _       _       _       start_char=85150|end_char=85153|ner=O
9       à       à       ADP     _       _       11      case    _       _
10      les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   11      det     _       _
11      gens    gens    NOUN    _       Number=Plur     8       obl:arg _       start_char=85154|end_char=85158|ner=O
```

C'est ainsi que le nom *gens* aurait dû être rattaché au verbe *donner* et non pas au nom *confiance*.

- Le nom du syntagme nominal n'est pas rattaché au verbe, mais à un autre mot qui aurait dû dépendre de la tête nominale, comme dans la capture suivante :

25)

```
# text = on nous donne plein de plein d' éléments pour euh pouvoir aborder les les les les enfants euh
# sent_id = cefc-crfp-PRI-BRI-1-290
1       on      on      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Ind    3       nsubj   _       start_char=101921|end_char=101923|ner=O
2       nous    il      PRON    _       Number=Plur|Person=1|PronType=Prs       3       nsubj   _       start_char=101924|end_char=101928|ner=O
3       donne   donner  VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin    0       root    _       start_char=101929|end_char=101934|ner=O
4       plein   plein   ADJ     _       Gender=Masc|Number=Sing 3       xcomp   _       start_char=101935|end_char=101940|ner=O
5       de      de      ADP     _       _       6       case    _       start_char=101941|end_char=101943|ner=O
6       plein   plein   NOUN    _       Gender=Masc|Number=Sing 4       obl:arg _       start_char=101944|end_char=101949|ner=O
7       d'      de      ADP     _       _       8       case    _       start_char=101950|end_char=101952|ner=O
8       éléments        élément NOUN    _       Gender=Masc|Number=Plur 6       nmod    _       start_char=101953|end_char=101961|ner=O
9       pour    pour    ADP     _       _       11      mark    _       start_char=101962|end_char=101966|ner=O
10      euh     lui     PRON    _       Number=Sing|Person=3|PronType=Prs       11      iobj    _       start_char=101967|end_char=101970|ner=O
11      pouvoir pouvoir VERB    _       VerbForm=Inf    3       advcl   _       start_char=101971|end_char=101978|ner=O
12      aborder aborder VERB    _       VerbForm=Inf    11      xcomp   _       start_char=101979|end_char=101986|ner=O
13      les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   17      det     _       start_char=101987|end_char=101990|ner=O
14      les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   17      det     _       start_char=101991|end_char=101994|ner=O
15      les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   17      det     _       start_char=101995|end_char=101998|ner=O
16      les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   17      det     _       start_char=101999|end_char=102002|ner=O
17      enfants enfant  NOUN    _       Number=Plur     12      obj     _       start_char=102003|end_char=102010|ner=O
18      euh     euh     X       _       _       17      appos   _       start_char=102011|end_char=102014|ner=O
```

Dans ce cas, le nom *éléments* aurait dû être rattaché au verbe et non pas à l'adjectif *plein*. En plus, il y a des erreurs au niveau des étiquettes morpho-syntaxiques, comme l'adjectif *plein* qui est étiqueté en tant qu'adjectif et en tant que nom.

Une source du bruit au-delà d'une mauvaise annotation au niveau des relations de dépendance est constituée par des erreurs d'omission dues à des problèmes de segmentation en phrase, comme :

26)

```
# text = mais quand j' embêterai les autres avec mes histoires qu' est-ce que ça donnera de plus ça donnera pas la guérison à m
# sent_id = cefc-crfp-PRI-AMI-3-547
1       mais    mais    CCONJ   _       _       10      cc      _       start_char=103814|end_char=103818|ner=O
2       quand   quand   SCONJ   _       _       4       mark    _       start_char=103819|end_char=103824|ner=O
3       j'      il      PRON    _       Number=Sing|Person=1|PronType=Prs       4       nsubj   _       start_char=103825|end_char=103827|ner=O
4       embêterai       embêter VERB    _       Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin    10      advcl   _       start_char=103828|end_char=103837|ner=O
5       les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   6       det     _       start_char=103838|end_char=103841|ner=O
6       autres  autre   PRON    _       Number=Plur|Person=3|PronType=Ind       4       obj     _       start_char=103842|end_char=103848|ner=O
7       avec    avec    ADP     _       _       9       case    _       start_char=103849|end_char=103853|ner=O
8       mes     son     DET     _       Number=Plur|Number[psor]=Sing|Person[psor]=1|Poss=Yes|PronType=Prs       9       det     _       start_char=103854|end_char=103857|ner=O
9       histoires       histoire        NOUN    _       Gender=Fem|Number=Plur  4       obl:mod _       start_char=103858|end_char=103867|ner=O
10      qu'     que     PRON    _       PronType=Int    0       root    _       start_char=103868|end_char=103871|ner=O
11      est     être    AUX     _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin    10      cop     _       start_char=103872|end_char=103875|ner=O
12      -ce     ce      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem    10      expl:subj       _       start_char=103875|end_char=103878|ner=O
13      que     que     PRON    _       PronType=Rel    15      obj     _       start_char=103879|end_char=103882|ner=O
14      ça      ça      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem    15      nsubj   _       start_char=103883|end_char=103885|ner=O
15      donnera donner  VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin    10      advcl:cleft     _       start_char=103886|end_char=103893|ner=O
16      de      de      ADP     _       _       19      advmod  _       start_char=103894|end_char=103896|ner=O
17      plus    plus    ADV     _       _       16      fixed   _       start_char=103897|end_char=103901|ner=O
18      ça      ça      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem    19      nsubj   _       start_char=103902|end_char=103904|ner=O
19      donnera donner  VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin    10      advcl:cleft     _       start_char=103905|end_char=103912|ner=O
20      pas     pas     ADV     _       Polarity=Neg    19      advmod  _       start_char=103913|end_char=103916|ner=O
21      la      le      DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art       22      det     _       start_char=103917|end_char=103919|ner=O
22      guérison        guérison        NOUN    _       Gender=Fem|Number=Sing  19      obj     _       start_char=103920|end_char=103928|ner=O
23      à       à       ADP     _       _       24      case    _       start_char=103929|end_char=103930|ner=O
24      m       mètre   NOUN    _       Gender=Masc|Number=Sing 22      nmod    _       start_char=103931|end_char=103932|ner=O
```

Dans cet exemple, le syntagme prépositionnel n'est pas complet et par conséquent, la phrase ne peut pas être pertinente dans le cadre d'une analyse du phénomène de l'alternance des compléments.

Puisque je ne souhaite pas renoncer à certaines données, j'ai décidé d'inclure le maximum de cas d'erreurs d'annotation à condition qu'ils soient récurrents dans le sous-corpus.

Les scores des mesures pour chaque type de structure sont présents dans le *Tableau 8* :

Type de structure	**Précision**	**Rappel**	**F-mesure**
donner + SN + SP	88%	100%	0.93
donner+SN(pron) + SP	-	-	-
CLI + donner + SN	99%	98%	0.98
CLI + donner + SN (pron)	-	-	-
CLI + donner + SP	100%	100%	1

CLI + donner + SN + SP(disloqué)	75%	100%	0.86
CLI+ donner + SN (pron) + SP(disloqué)	-	-	-
CLI1 + CLI2 + donner + SN(disloqué)	-	-	-
CLI1 + CLI2 + donner + SN(pron, disloqué)	-	-	-
CLI1 + CLI2 + donner + SP(disloqué)	-	-	-
CLI1 + CLI2 + donner	100%	100%	1

Tableau 8 *Résultats des mesures dans l'évaluation en fonction de chaque sous-type*

Vu les résultats des mesures calculées, les scores sont assez hauts pour tous les types de structures identifiées. En ce qui concerne la structure de type CLI + donner + SN + SP(disloqué) la précision est assez faible (75%), fait qui influence la F-mesure (0.86)

Puisque cette évaluation ne permet pas d'observer le rendement global de l'extraction, une vérification manuelle du jeu de données final est requise afin d'assurer des résultats fiables à la suite des analyses statistiques.

La sous-section suivante présente le jeu de données final obtenu après avoir appliqué la chaîne de traitement mise en place pour le corpus CEFC entier sélectionné.

4.4 Du sous-corpus *CRFP* au corpus *CEFC*

Puisque la méthode choisie a donné de bons résultats à la suite de l'évaluation, j'ai choisi d'implémenter la même démarche pour tout le corpus sélectionné (3.2). Comme la méthode d'évaluation utilisée est de type intrinsèque et il n'est pas possible de voir en détail la qualité de l'extraction finale, une vérification manuelle est requise. Cette vérification manuelle offre des indices importants sur les éventuelles limites restantes à la suite de l'utilisation d'une telle méthode d'extraction.

Les résultats de l'extraction des phrases de type A sont visibles dans le *Tableau 9* :

Type de structure	Nombre de phrases	% du nombre de phrases contenant le verbe donner du corpus CEFC	% du nombre de phrases contenant le verbe donner du corpus CEFC [+type A]
donner + SN + SP	117	3.9%	95.12%
donner + SN(pron) + SP	6	0.2%	4.88%

Tableau 9 *Les résultats chiffrés des données de type A le corpus CEFC*

Selon les résultats, tout comme pour le sous-corpus *CRFP*, la structure de type donner + SN + SP est beaucoup plus fréquente que la structure de type donner + SN(pron) + SP (95.12% vs. 4.88%). Cependant, le système a pu extraire 6 phrases de type donner + SN(pron) + SP contrairement au cas du corpus *CRFP* où le système n'a trouvé aucune phrase contenant cette structure.

Au niveau du grand type A, il existe 123 phrases dans lesquelles les arguments du verbe *donner* sont présents strictement dans la zone postverbale, soit 4.1% du nombre des phrases contenant le lemme *donner* et 0.33% du nombre des phrases du corpus sélectionné.

Quant aux phrases dans lesquelles les arguments du verbe *donner* se retrouvent autant dans la zone préverbale que dans la zone postverbale (type B), les résultats sont synthétisés dans le *Tableau 10* :

Type de structure	Nombre de phrases	% du nombre de phrases contenant le verbe donner du corpus CEFC	% du nombre de phrases contenant le verbe donner du corpus CEFC [+ type B]
CLI + donner + SN	742	26.57%	92.29%
CLI + donner + SN(pron)	25	0.80%	3.11%
CLI + donner + SP	16	0.47%	1.99%
CLI + donner + SN + SP(disloqué)	11	0.37%	1.37%
CLI + donner + SN(pron) + SP(disloqué)	0	0%	0%

CLI1 + CLI2 + donner + SN(disloqué)	9	0.27%	1.12%
CLI1 + CLI2 + donner + SN(pron, disloqué)	1	0.03%	0.13%
CLI1 + CLI2 + donner + SP(disloqué)	0	0%	0%

Tableau 10 *Les résultats chiffrés des données de type B le corpus CEFC*

Tout comme dans le cas du corpus CEFC, la structure la plus fréquente est CLI + donner + SN qui domine les autres structures (92.29%) et se retrouve en même temps dans une partie considérable des phrases contenant le lemme *donner* du corpus *CEFC* (26.57%). Contrairement au sous-corpus *CRFC* où la structure de type CLI + donner + SN(pron) n'a pas été retrouvée, cette fois-ci elle représente la deuxième la plus fréquente structure de type B (3.11%) suivie par la structure de type CLI + donner + SP (1.99%).

Les structures de type CLI + donner + SN(pron) + SP(disloqué) et CLI1 + CLI2 + donner + SP(disloqué) n'apparaissent toujours pas dans le grand corpus sélectionné, tandis que deux structures émergent, notamment CLI1 + CLI2 + donner + SN(disloqué) (1.12%) et CLI1 + CLI2 + donner + SN(pron, disloqué) (0.13%).

Au niveau des phrases contenant le lemme *donner* du corpus CEFC, ces structures ne dépassent pas en général le seuil de 1%, l'exception étant la structure de type CLI + donner + SN.

En ce qui concerne les phrases dans lesquelles les arguments du verbe *donner* se retrouvent strictement dans la zone préverbale (type C), les résultats sont synthétisés dans le *Tableau 11* :

Type de structure	Nombre de phrases	% du nombre de phrases contenant le verbe donner du corpus CEFC	% du nombre de phrases contenant le verbe donner du corpus CEFC [+type C]
CLI1 + CLI2 + donner	35	1.10%	100%

Tableau 11 *Les résultats chiffrés des données de type C dans le corpus CEFC*

D'après les résultats, la seule structure ciblée du type C se retrouve en proportion de 1.10% dans les phrases contenant le verbe *donner* du corpus *CEFC* entier sélectionné.

Au niveau de chaque grand type de structure, la répartition dans le corpus *CEFC* (constitué par les phrases contenant le lemme *donner* du sous-corpus *CRFP*) est visible dans le *Tableau 12* :

Grand type de structure	% du nombre des phrases contenant le verbe donner dans le corpus CEFC
Type A	4.1%
Type B	26.77%
Type C	1.17%
Autres phrases	67.96%

Tableau 12 *La répartition des données selon le type de structure dans le corpus CEFC*

Parmi les 1018 phrases que le script a extraites indépendamment du grand type de structure, il est facilement observable que les tendances déjà observées dans le sous-corpus *CRFP* se maintiennent aussi pour le corpus CEFC. Tout comme dans le cas de l'échantillon étudié, la plupart des phrases du corpus CEFC contenant le lemme *donner* ne présentent aucun des grands types cibles (67.96%). Le grand type le plus fréquent est le type B (26.77%), suivi par le type A (4.1%) et le type C (1.17%). Au sens large, les structures appartenant au grand type B et au grand type C (prises ensemble) sont plus fréquentes que ceux qui ne contiennent pas de pronoms clitiques (comme dans le cas du sous-corpus *CRFP*).

La sous-section suivante présente les limites qui ont été dépassées grâce à la méthodologie mise en place, mais aussi les limites qui demeurent et ne peuvent être dépassées que par une révision manuelle du jeu de données afin de permettre des analyses pertinentes qui ne soient pas biaisées par les données.

4.5 Améliorations réalisées et limites

Dans cette sous-section, je mettrai d'abord en discussion les limites qui ont été dépassées grâce à la démarche mise en place et dans un deuxième temps les limites qui demeurent malgré les résultats remarquables de l'évaluation à la suite de l'extraction.

A. Limites dépassées

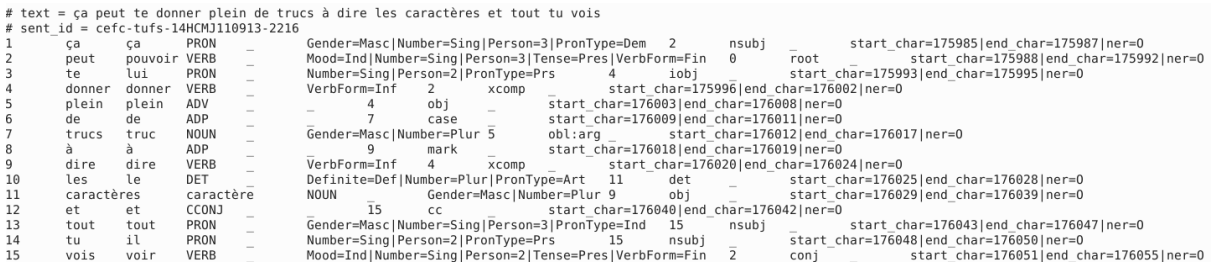
La première limite importante qui a été dépassée grâce à l'utilisation d'un script au détriment de l'extraction directement depuis la plateforme Orfeo a été le fait de pouvoir extraire des résultats de tous les sous-corpus à la fois, fait qui n'est pas possible dans le cas de la plateforme.

Une autre limite a été l'annotation du corpus qui ne permettait pas une extraction efficace des données et laissait de côté la majorité des phrases qui avaient le potentiel d'être pertinentes pour mon étude. Parmi les soucis récurrents, il y avait le déterminant *des* marqué en tant que préposition et les dépendants mal tracés au niveau des relations de dépendance (voir section 3.4). Grâce à la conversion du format orfeo sous format CoNNL-U, le souci du déterminant *des* n'existe plus et les dépendants ne représentent plus un problème insurmontable avec un script informatique.

Cependant, il y a toujours des erreurs d'annotation, comme :

- La tête du SN du verbe dépend d'un autre mot qui devrait dépendre de lui et qui à son tour dépend du verbe

27)



```
# text = ça peut te donner plein de trucs à dire les caractères et tout tu vois
# sent_id = cefc-tufs-14HCMJ110913-2216
1       ça      ça      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem    2       nsubj   _       start_char=175985|end_char=175987|ner=O
2       peut    pouvoir VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root    _       start_char=175988|end_char=175992|ner=O
3       te      lui     PRON    _       Number=Sing|Person=2|PronType=Prs       4       iobj    _       start_char=175993|end_char=175995|ner=O
4       donner  donner  VERB    _       VerbForm=Inf    2       xcomp   _       start_char=175996|end_char=176002|ner=O
5       plein   plein   ADV     _       _       4       obj     _       start_char=176003|end_char=176008|ner=O
6       de      de      ADP     _       _       7       case    _       start_char=176009|end_char=176011|ner=O
7       trucs   truc    NOUN    _       Gender=Masc|Number=Plur 5       obl:arg _       start_char=176012|end_char=176017|ner=O
8       à       à       ADP     _       _       9       mark    _       start_char=176018|end_char=176019|ner=O
9       dire    dire    VERB    _       VerbForm=Inf    4       xcomp   _       start_char=176020|end_char=176024|ner=O
10      les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   11      det     _       start_char=176025|end_char=176028|ner=O
11      caractères      caractère       NOUN    _       Gender=Masc|Number=Plur 9       obj     _       start_char=176029|end_char=176039|ner=O
12      et      et      CCONJ   _       _       15      cc      _       start_char=176040|end_char=176042|ner=O
13      tout    tout    PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Ind    15      nsubj   _       start_char=176043|end_char=176047|ner=O
14      tu      il      PRON    _       Number=Sing|Person=2|PronType=Prs       15      nsubj   _       start_char=176048|end_char=176050|ner=O
15      vois    voir    VERB    _       Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin   2       conj    _       start_char=176051|end_char=176055|ner=O
```

Cet exemple montre comment la tête du SN *trucs* dépend de l'adjectif plein qui est à son tour rattaché au verbe.

- La tête du SP dépend d'un nom à l'intérieur du SP qui dépend à son tour d'un autre mot étant rattaché au verbe.

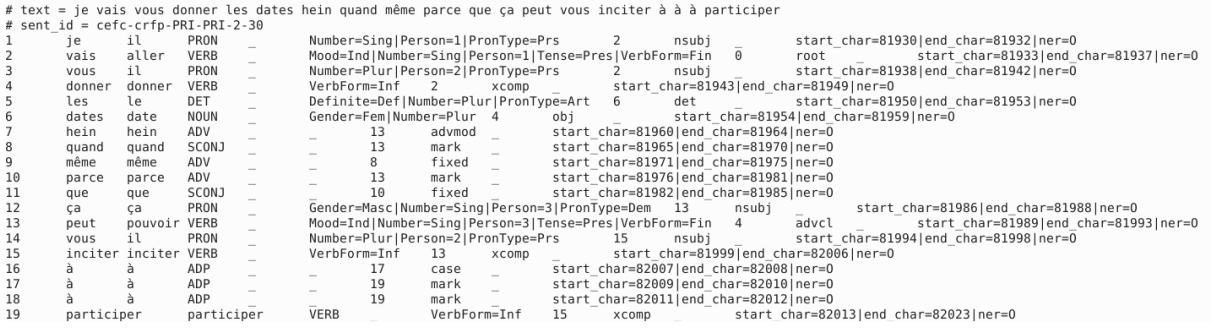
28)

```
# text = je au départ je donne quand même une une une direction euh au personnel au sol en disant comme ça bon va tel end-~
# sent_id = cefc-crfp-PRI-LIM-1-319
1       je      il      PRON    _       Number=Sing|Person=1|PronType=Prs       6       nsubj   _       start_char=104797|end_char=104799|ner=O
2-3     au      _       _       _       _       _       _       start_char=104800|end_char=104802|ner=O
2       à       à       ADP     _       _       4       case    _       _
3       le      le      DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       4       det     _       _
4       départ  départ  NOUN    _       Gender=Masc|Number=Sing 6       obl:mod _       start_char=104803|end_char=104809|ner=O
5       je      il      PRON    _       Number=Sing|Person=1|PronType=Prs       6       nsubj   _       start_char=104810|end_char=104812|ner=O
6       donne   donner  VERB    _       Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin   0       root    _       start_char=104813|end_char=104818|ner=O
7       quand   quand   SCONJ   _       _       6       advmod  _       start_char=104819|end_char=104824|ner=O
8       même    même    ADV     _       _       7       fixed   _       start_char=104825|end_char=104829|ner=O
9       une     un      DET     _       Definite=Ind|Gender=Fem|Number=Sing|PronType=Art        10      det     _       start_char=104830|end_char=104833|ner=O
10      une     un      PRON    _       Gender=Fem|Number=Sing|Person=3|PronType=Ind     6       obj     _       start_char=104834|end_char=104837|ner=O
11      une     un      DET     _       Definite=Ind|Gender=Fem|Number=Sing|PronType=Art        12      det     _       start_char=104838|end_char=104841|ner=O
12      direction       direction       NOUN    _       Gender=Fem|Number=Sing  6       obj     _       start_char=104842|end_char=104851|ner=O
13      euh     euh     ADV     _       _       12      advmod  _       start_char=104852|end_char=104855|ner=O
14-15   au      _       _       _       _       _       _       start_char=104856|end_char=104858|ner=O
14      à       à       ADP     _       _       16      case    _       _
15      le      le      DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       16      det     _       _
16      personnel       personnel       NOUN    _       Gender=Masc|Number=Sing 12      nmod    _       start_char=104859|end_char=104868|ner=O
17-18   au      _       _       _       _       _       _       start_char=104869|end_char=104871|ner=O
17      à       à       ADP     _       _       19      case    _       _
18      le      le      DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       19      det     _       _
19      sol     sol     NOUN    _       Gender=Masc|Number=Sing 16      nmod    _       start_char=104872|end_char=104875|ner=O
20      en      en      ADP     _       _       21      mark    _       start_char=104876|end_char=104878|ner=O
21      disant  dire    VERB    _       Tense=Pres|VerbForm=Part        6       advcl   _       start_char=104879|end_char=104885|ner=O
22      comme   comme   SCONJ   _       _       25      mark    _       start_char=104886|end_char=104891|ner=O
23      ça      ça      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem    25      nsubj   _       start_char=104892|end_char=104894|ner=O
24      bon     bon     ADJ     _       Gender=Masc|Number=Sing 25      xcomp   _       start_char=104895|end_char=104898|ner=O
25      va      aller   VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   21      advcl   _       start_char=104899|end_char=104901|ner=O
26      tel     tel     DET     _       Gender=Masc|Number=Sing|PronType=Ind     27      det     _       start_char=104902|end_char=104905|ner=O
27      end-~   end-~   NOUN    _       Gender=Masc|Number=Sing 25      nsubj   _       start_char=104906|end_char=104911|ner=O
```

Dans cette phrase, la préposition *à* dépend du nom *personnel* qui aurait dû dépendre à son tour directement du verbe, mais il dépend de la tête du SN *direction* qui est rattaché finalement au verbe *donner*.

- Le pronom clitique dépend des verbes *aller* et *pouvoir* dans le cas de l'emploi du futur proche ou d'un conditionnel réalisé avec le verbe *pouvoir*

29)



```
# text = je vais vous donner les dates hein quand même parce que ça peut vous inciter à à à participer
# sent_id = cefc-crfp-PRI-PRI-2-30
1       je      il      PRON    _       Number=Sing|Person=1|PronType=Prs       2       nsubj   _       start_char=81930|end_char=81932|ner=O
2       vais    aller   VERB    _       Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin   0       root    _       start_char=81933|end_char=81937|ner=O
3       vous    il      PRON    _       Number=Plur|Person=2|PronType=Prs       2       nsubj   _       start_char=81938|end_char=81942|ner=O
4       donner  donner  VERB    _       VerbForm=Inf    2       xcomp   _       start_char=81943|end_char=81949|ner=O
5       les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   6       det     _       start_char=81950|end_char=81953|ner=O
6       dates   date    NOUN    _       Gender=Fem|Number=Plur  4       obj     _       start_char=81954|end_char=81959|ner=O
7       hein    hein    ADV     _       _       13      advmod  _       start_char=81960|end_char=81964|ner=O
8       quand   quand   SCONJ   _       _       13      mark    _       start_char=81965|end_char=81970|ner=O
9       même    même    ADV     _       _       8       fixed   _       start_char=81971|end_char=81975|ner=O
10      parce   parce   ADV     _       _       13      mark    _       start_char=81976|end_char=81981|ner=O
11      que     que     SCONJ   _       _       10      fixed   _       start_char=81982|end_char=81985|ner=O
12      ça      ça      PRON    _       Gender=Masc|Number=Sing|Person=3|PronType=Dem    13      nsubj   _       start_char=81986|end_char=81988|ner=O
13      peut    pouvoir VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   4       advcl   _       start_char=81989|end_char=81993|ner=O
14      vous    il      PRON    _       Number=Plur|Person=2|PronType=Prs       15      nsubj   _       start_char=81994|end_char=81998|ner=O
15      inciter inciter VERB    _       VerbForm=Inf    13      xcomp   _       start_char=81999|end_char=82006|ner=O
16      à       à       ADP     _       _       17      case    _       start_char=82007|end_char=82008|ner=O
17      à       à       ADP     _       _       19      mark    _       start_char=82009|end_char=82010|ner=O
18      à       à       ADP     _       _       19      mark    _       start_char=82011|end_char=82012|ner=O
19      participer      participer      VERB    _       VerbForm=Inf    15      xcomp   _       start_char=82013|end_char=82023|ner=O
```

Dans cette phrase, le pronom clitique *vous* dépend de l'auxiliaire avec lequel se construit le futur proche du verbe *donner* au lieu de dépendre directement du lemme *donner*.

D'autres limites observées au niveau de tous les sous-corpus après la vérification manuelle ont été les cas des phrases qui se répètent, ce qui peut engendrer des soucis. Elles ont été dépassées par l'élimination de tous les doublons.

Un autre problème provient des mots qui se répètent à cause de la présence des disfluences dans la transcription du langage spontané, comme :

30)

```
# text = et avant de les donner aux aux propriétaires ou aux locataires et caetera avaient imaginé de les mettre à la disposition du public dans un temps
# sent_id = cefc-valibel-norHJ1r-1484
1       et      et      CCONJ   _       _       18      cc      _       start_char=393839|end_char=393841|ner=O
2       avant   avant   ADV     _       _       18      advmod  _       start_char=393842|end_char=393847|ner=O
3       de      de      ADP     _       _       5       mark    _       start_char=393848|end_char=393850|ner=O
4       les     le      PRON    _       Number=Plur|Person=3|PronType=Prs       5       obj     _       start_char=393851|end_char=393854|ner=O
5       donner  donner  VERB    _       VerbForm=Inf    2       ccomp   _       start_char=393855|end_char=393861|ner=O
6-7     aux     _       _       _       _       _       _       start_char=393862|end_char=393865|ner=O
6       à       à       ADP     _       _       10      case    _       _
7       les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   10      det     _       _
8-9     aux     _       _       _       _       _       _       start_char=393866|end_char=393869|ner=O
8       à       à       ADP     _       _       10      case    _       _
9       les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   10      det     _       _
10      propriétaires   propriétaire    NOUN    _       Gender=Masc|Number=Plur 5       obl:arg _       start_char=393870|end_char=393883|ner=O
11      ou      ou      CCONJ   _       _       14      cc      _       start_char=393884|end_char=393886|ner=O
12-13   aux     _       _       _       _       _       _       start_char=393887|end_char=393890|ner=O
12      à       à       ADP     _       _       14      case    _       _
13      les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   14      det     _       _
14      locataires      locataire       NOUN    _       Gender=Masc|Number=Plur 10      conj    _       start_char=393891|end_char=393901|ner=O
15      et      et      CCONJ   _       _       18      cc      _       start_char=393902|end_char=393904|ner=O
16      caetera caetera PROPN   _       _       18      nsubj   _       start_char=393905|end_char=393912|ner=O
17      avaient avoir   AUX     _       Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin    18      aux:tense       _       start_char=393913|end_char=393920|ner=O
18      imaginé imaginer        VERB    _       Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part        0       root    _       start_char=393921|end_char=393928|ner=O
19      de      de      ADP     _       _       21      mark    _       start_char=393929|end_char=393931|ner=O
20      les     le      PRON    _       Number=Plur|Person=3|PronType=Prs       21      obj     _       start_char=393932|end_char=393935|ner=O
21      mettre  mettre  VERB    _       VerbForm=Inf    18      xcomp   _       start_char=393936|end_char=393942|ner=O
22      à       à       ADP     _       _       24      case    _       start_char=393943|end_char=393944|ner=O
23      la      le      DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        24      det     _       start_char=393945|end_char=393947|ner=O
24      disposition     disposition     NOUN    _       Gender=Fem|Number=Sing  21      obl:arg _       start_char=393948|end_char=393959|ner=O
25-26   du      _       _       _       _       _       _       start_char=393960|end_char=393962|ner=O
25      de      de      ADP     _       _       27      case    _       _
26      le      le      DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       27      det     _       _
27      public  public  NOUN    _       Gender=Masc|Number=Sing 24      nmod    _       start_char=393963|end_char=393969|ner=O
28      dans    dans    ADP     _       _       30      case    _       start_char=393970|end_char=393974|ner=O
29      un      un      DET     _       Definite=Ind|Gender=Masc|Number=Sing|PronType=Art       30      det     _       start_char=393975|end_char=393977|ner=O
30      temps   temps   NOUN    _       Gender=Masc|Number=Sing 21      obl:mod         start_char=393978|end_char=393983|ner=O
```

Cet exemple illustre très bien les hésitations qui peuvent apparaître dans le langage oral spontané qui ont été transcrites, mais qui rendent la tâche d'extraction des données plus difficile.

Malgré ces problèmes, toutes ces erreurs peuvent être gérées à l'aide des conditions supplémentaires au niveau de l'extracteur, comme le fait d'inclure parmi les phrases extraites qui ne contiennent pas d'erreurs d'annotation, les cas de dépendance syntaxique présents dans les exemples (27), (28) et (29).

B. Limites restantes

Même si l'analyse morpho-syntaxique réalisée par le parser *Stanza* (Qi et al., 2020) est meilleure que celle réalisée par le parser *MACAON* (Nasr et al., 2011), il reste encore des cas isolés qui ne se soumettent à aucune tendance générale et une fois ajoutées les conditions qui permettraient de les extraire, elles généreraient plus d'inconvénients que d'avantages, notamment une quantité énorme de bruit difficile à gérer manuellement.

Les hésitations des locuteurs qui ont été transcrites telles quelles (exemple - 30) peuvent impliquer des problèmes au niveau des relations de dépendance pour l'analyseur syntaxique et plus tard d'annotation, car la structure-cible est difficilement identifiable pour un système automatique. D'ailleurs, même si dans le cas du parser *Stanza* l'étiquetage est bien meilleur que dans le cas du parser *MACAON* (Nasr et al., 2011), il peut y avoir des cas où le système se trompe (exemples - 27, 28, 29).

Un autre problème qui peut apparaître est l'existence de plusieurs syntagmes à la suite des disfluences qui interviennent dans le langage oral, comme dans l'exemple qui suit :

31)

```
# text = voilà absolument nous en nous nous demandons aux enfants qu' ils donnent une leçon à leurs parents aux adultes
# sent_id = cefc-valibel-jtaDC1r-424
1       voilà   voilà   VERB    _       _       0       root    _       start_char=421310|end_char=421315|ner=O
2       absolument      absolument      ADV     _       _       1       advmod  _       start_char=421316|end_char=421326|ner=O
3       nous    il      PRON    _       Number=Plur|Person=1|PronType=Prs       7       nsubj   _       start_char=421327|end_char=421331|ner=O
4       en      en      PRON    _       Person=3|PronType=Prs   7       iobj    _       start_char=421332|end_char=421334|ner=O
5       nous    il      PRON    _       Number=Plur|Person=1|PronType=Prs       7       nsubj   _       start_char=421335|end_char=421339|ner=O
6       nous    il      PRON    _       Number=Plur|Person=1|PronType=Prs       7       nsubj   _       start_char=421340|end_char=421344|ner=O
7       demandons       demander        VERB    _       Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin   1       ccomp   _       start_char=421345|end_char=421354|ner=O
8-9     aux     _       _       _       _       _       _       _       start_char=421355|end_char=421358|ner=O
8       à       à       ADP     _       _       10      case    _       _
9       les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   10      det     _       _
10      enfants enfant  NOUN    _       Number=Plur     7       obl:arg _       start_char=421359|end_char=421366|ner=O
11      qu'     que     SCONJ   _       _       13      mark    _       start_char=421367|end_char=421370|ner=O
12      ils     il      PRON    _       Gender=Masc|Number=Plur|Person=3|PronType=Prs   13      nsubj   _       start_char=421371|end_char=421374|ner=O
13      donnent donner  VERB    _       Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   7       ccomp   _       start_char=421375|end_char=421382|ner=O
14      une     un      DET     _       Definite=Ind|Gender=Fem|Number=Sing|PronType=Art       15      det     _       start_char=421383|end_char=421386|ner=O
15      leçon   leçon   NOUN    _       Gender=Fem|Number=Sing  13      obj     _       start_char=421387|end_char=421392|ner=O
16      à       à       ADP     _       _       18      case    _       start_char=421393|end_char=421394|ner=O
17      leurs   son     DET     _       Number=Plur|Number[psor]=Plur|Person[psor]=3|Poss=Yes|PronType=Prs      18      det     _       start_char=421395|end_char=421400|ner=O
18      parents parent  NOUN    _       Gender=Masc|Number=Plur 13      obl:arg _       start_char=421401|end_char=421408|ner=O
19-20   aux     _       _       _       _       _       _       _       start_char=421409|end_char=421412|ner=O
19      à       à       ADP     _       _       21      case    _       _
20      les     le      DET     _       Definite=Def|Number=Plur|PronType=Art   21      det     _       _
21      adultes adulte  NOUN    _       Gender=Masc|Number=Plur 13      obl:arg _       start_char=421413|end_char=421420|ner=O
```

Dans cette phrase, puisque le locuteur s'est auto-corrigé, il existe deux syntagmes prépositionnels *à leurs parents* et *aux adultes*, ce qui rend plus difficile le processus d'extraction et plus tard d'annotation automatique. Afin d'éviter d'autres soucis de bruit, j'ai choisi de gérer à la main les doublons qui peuvent apparaître en sortie à cause d'une double identification du SP au niveau de la phrase extraite et j'ai décidé de prendre en considération la deuxième forme mentionne (l'auto-correction) car elle représente le contenu auquel le locuteur veut finalement se référer.

Au-delà des limites qui peuvent être facilement identifiées à l'aide du gold standard réalisé sur le sous-corpus *CRFP*, il peut encore exister des cas de silence qu'il serait impossible d'identifier dans les cas des autres sous-corpus.

Afin d'assurer une très bonne qualité au niveau de l'extraction des données, il faut vérifier manuellement à la fin tout le jeu de données créé, tâche qui est beaucoup moins sollicitante dans le cas d'un extracteur bien construit que dans le cas d'un travail manuel sur environ 4 millions de mots.

La section suivante décrit la deuxième grande étape de mon travail méthodologique, notamment l'annotation qui prépare la dernière étape, notamment l'analyse de l'alternance d'ordre des compléments en français oral spontané.

5 Annotation des données pour l'étude de l'alternance des compléments du verbe donner

L'annotation des données est une étape obligatoire dans la réalisation des analyses statistiques sur des phénomènes syntaxiques, comme dans le cas de l'ordre des compléments postverbaux en français oral spontané. Il est important d'annoter les données afin de tester l'impact des facteurs étudiés. Dans cette section, je présenterai les variables annotées pour observer les facteurs étudiés et comment j'ai réalisé l'annotation du jeu de données obtenu de façon semi-automatique (section 5.1) et de façon automatique en fonction des variables annotées (section 5.2).

5.1 Annotation semi-automatique

Il existe des facteurs comme l'animéité, l'accessibilité discursive et le type de verbe (support ou pas support) qui ne permettent pas une annotation exclusivement automatique.

Afin de rendre le processus plus efficace et ne pas recourir à une annotation manuelle, j'ai réalisé une annotation semi-automatique à l'aide la plateforme d'annotation *INCEpTION*[8] (Klie et al., 2018). Cette plateforme donne la possibilité de créer des couches d'annotation en fonction des besoins du projet et d'associer les étiquettes souhaitées aux couches créées.

Les variables que j'ai annotées semi-automatiquement à l'aide de la plateforme d'annotation *INCEpTION* et qui représente chacune une couche d'annotation sont les suivantes :

- *SN_animéité*

Cette variable indique si le référent du SN est animé ou ne l'est pas. Elle peut avoir deux valeurs qui génèrent deux étiquettes : *SN_animé*, si le référent est animé et *SN_inanimé*, si le référent ne l'est pas.

L'annotation de cette variable s'est avérée assez difficile car il y avait des cas où le statut du SN n'était pas évident contrairement a des cas ou il n'y avait pas de doutes, comme dans l'exemple 32) :

32) On donne [des prix]$_{\text{SNinanimé}}$ aux Suisses.

Les cas qui ont posé des problèmes ont été représentés par les noms qui renvoient à des êtres. Afin de garder une cohérence dans l'annotation j'ai choisi de considérer ces noms en tant que des SN inanimés, comme le cas du nom *naissance* present dans l'exemple 33) :

33) Et l' histoire aussi c' est qu'il y a un chameau qui a de la peine à donner [naissance]$_{\text{SNinanimé}}$ à son enfant quoi.

J'ai choisi d'annoter cette variable, car le facteur animéité a fait l'objet de plusieurs études et il semble avoir une influence dans le choix de l'ordre des compléments (Thuilier et al., 2021).

- *SP_animéité*

Cette variable indique si le référent du SP est animé ou il ne l'est pas. Elle peut avoir deux valeurs qui génèrent deux étiquettes : *SP_animé*, si le référent est animé et *SP_inanimé*, si le référent ne l'est pas.

[8] https://inception-project.github.io/

Il y avait des cas difficiles pour cette variable aussi, comme les référents qui renvoient à des êtres humains. Afin de garder une cohérence dans l'annotation, j'ai considéré ces cas difficiles comme des syntagmes dont le référent est inanimé, comme dans l'exemple :

34) Voilà par contre je donne dix mille francs par année [au Secrétariat national]$_{\text{SPinanimé}}$ pour qu'il réfléchisse puis qu' il me trouve des projets à faire voilà.

Tout comme pour le SN, cette variable a la même importance, mais du côté du SP.

- *SN_accessibilité*

Cette variable indique si le référent du SN est accessible ou il ne l'est pas. Elle peut avoir trois valeurs qui génèrent trois étiquettes : *SN_accessible*, si le référent est accessible, *SN_inaccessible*, si le référent ne l'est pas et *SN_inferable*, si le référent est déduit pas raisonnement à partir des référents déjà évoqués (Prince, 1981).

Les choix dans l'annotation de cette variable n'ont pas été toujours évidents, car il y avait des cas où on aurait pu hésiter dans l'annotation entre les syntagmes dont le référent aurait pu être considéré comme inférable ou bien inaccessible. C'est pour cette raison que j'ai décidé de considérer le référent comme inférable s'il est possible de faire un lien sémantique avec l'un des éléments présents dans la phrase antérieure extraite, comme dans l'exemple suivant :

35) Il y a pas de concours dedans. Mais juste on donne [des prix]$_{\text{SNinférable}}$ à ceux qui chantent bien.

Ainsi, le terme *concours* conduit à l'idée de l'existence des prix et on peut déduire la liaison entre les deux référents.

J'ai choisi d'annoter cette variable, car le facteur accessibilité discursive a fait l'objet des plusieurs études et d'après Faghiri & Thuilier (2018), il a une influence dans le choix de l'ordre des compléments.

- *SP_accessibilité*

Cette variable indique si le référent du SN est accessible ou il ne l'est pas. Elle peut avoir trois valeurs qui génèrent trois étiquettes : *SP_accessible*, si le référent est accessible, *SP_inaccessible*, si le référent ne l'est pas et *SP_inférable*, si le référent est déduit par raisonnement à partir des référents déjà évoqués (Prince, 1981).

Tout comme dans le cas de l'annotation de la variable *SN_accessibilité*, il y avait le même problème en ce qui concerne le doute entre le statut inferable ou inaccessible d'un référent qui se retrouve dans un lien logique avec un autre référent présent dans la phrase antérieure.

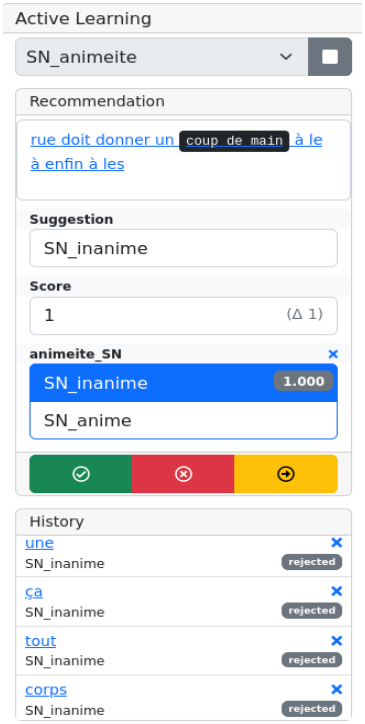
36) Mais dans un bus je veux dire un bus euh. C' est le principe du truc de donner de l' argent [au chauffeur]$_{\text{SPinférable}}$.

Ainsi, le terme *bus* implique l'idée de l'existence d'un chauffeur et on peut déduire la liaison entre les deux référents.

La plateforme *INCEpTION* permet une annotation semi-automatique, car elle contient des outils d'apprentissage automatique appelés *Recommenders* qui offrent des suggestions pendant le processus d'annotation en se basant sur la partie du texte qui a été déjà annoté préalablement. Plus la partie de texte annoté manuellement est grande, plus le système sera entraîné et il va offrir de meilleures suggestions lors du processus d'annotation. Les suggestions peuvent être acceptées, refusées ou ignorées. La technologie de l'outil qui est disponible et qui me sert dans mon annotation est celle d'un

string matcher. Cela veut dire que l'outil fera des suggestions et proposera les mêmes étiquettes pour des mots similaires aux mots qui ont déjà été annotés, comme on le voit dans l'exemple 32) :

37)



Dans ce cas, l'outil a recommandé l'étiquette *SN_inanime* pour la séquence *coup de main* et cette suggestion reçoit un score de correspondance maximum (1.000) pour le choix d'un syntagme nominal inanimé au détriment d'un syntagme nominal animé. Comme la séquence *donner un coup de main à quelqu'un* apparaît plusieurs fois dans les données et elle a déjà été annotée avec l'étiquette *SN_inanime*, il est normal que l'outil donne cette suggestion. L'outil propose également un historique de toutes les suggestions faites et chaque décision de l'annotateur. Il propose également des mesures pour évaluer la qualité des suggestions faites par le système dans le cas de chaque couche.

F1	Accuracy	Précision	Rappel
0.51	0.35	1	0.35

Tableau 13 Les scores des mesures pour la couche SN_animeite

Dans le cas de la couche *SN_animeite* qui indique un syntagme nominal dont le référent est animé ou inanimé, le score de la précision est maximum, car il y a des référents qui apparaissent plusieurs fois dans les données. En même temps, le score de l'accuracy (indique le rapport entre les étiquettes correctement détectées par le système et les étiquettes correctement ignorées par le système) est assez bas car la technologie basée sur l'identification d'une chaîne de caractère ne suffit pas pour identifier la bonne structure dans le bon endroit de la phrase. Le score du rappel est également assez bas vu le nombre diminué de cas où le système a identifié le bon référent dans le bon contexte. Comme la précision est très haute, le score de la mesure F1 (la moyenne harmonique entre la précision et le rappel) s'élève à 0.53.

F1	Accuracy	Précision	Rappel
0.53	0.44	1	0.36

Tableau 14 Les scores des mesures pour la couche SP_animeite

Tout comme dans le cas de la couche *SN_animeite*, les scores pour la couche *SP_animeite* sont assez similaires. Cependant, le score de l'accuracy est un peu plus haut cette fois-ci, ce qui signifie que le système arrive mieux à identifier les référents animés et inanimés des syntagmes prépositionnels.

F1	Accuracy	Précision	Rappel
0.32	0.36	0.5	0.23

Tableau 15 Les scores des mesures pour la couche SN_accessibilite_discursive

Dans le cas de la couche *SN_accessibilite_discursive* qui indique si le SN est accessible, inferable ou inaccessible, les scores diminuent encore plus surtout au niveau de la précision (0.5). Il est plus difficile pour le système de proposer de bonnes suggestions, car même s'il s'agit du même référent et presque le même contexte dans la phrase, il y a peu de chances qu'il fasse partie de la même structure informationnelle.

F1	Accuracy	Précision	Rappel
0.07	0.07	0.33	0.04

Tableau 16 *Les scores des mesures pour la couche SP_accessibilite_discursive*

Les scores pour la couche *SP_accessibilite_discursive* sont encore plus bas que les scores de la couche *SN_accessibilite_discursive* d'où la difficulté du système qui se base sur la correspondance au niveau des chaînes de caractère d'identifier la structure informationnelle de la phrase.

De façon générale, les scores des mesures d'évaluation pour la qualité des suggestions faites par l'outil sont assez réduits car la technologie d'un *string matcher* ne suffit pas dans l'identification de la bonne chaîne de caractères dans le bon contexte, mais cette démarche permet une annotation plus efficace au niveau du temps par rapport à une annotation exclusivement manuelle.

5.2 Annotation automatique

Afin de rendre plus efficace le processus d'annotation, j'ai annoté automatiquement la majorité des variables dont j'ai eu besoin dans mes analyses statistiques. Les variables que j'ai annotées de manière automatique à l'aide d'un script Python sont les suivantes :

- *phrase*

Cette variable indique la phrase pertinente qui a été extraite par le script informatique.

- *phrase_ant*

Cette variable représente la phrase antérieure à la phrase pertinente qui a été extraite par le script informatique. Cette variable est utile pour l'annotation de l'accessibilité discursive (Faghiri & Thuilier, 2018).

- *SN*

Cette variable indique le contenu du syntagme nominal qui appartient à la phrase pertinente.

- *SP*

Cette variable indique le contenu du syntagme prépositionnel qui appartient à la phrase pertinente.

- Longueur

Cette variable indique le type de SN (court ou long) et SP (court ou long) conformément à l'ordre d'apparition. Les valeurs possibles de la variable *Longueur* sont :

- *court long* : S'il y a un ordre de type SN-SP dans la phrase et la longueur en nombre de mots du SN est inférieure à la longueur en nombre de mots de la variable SP;
- *long-court* : S'il y a un ordre de type SN-SP dans la phrase et la longueur en nombre de mots du SN est supérieure à la longueur en nombre de mots du SP;
- *autre* : Les longueurs des deux syntagmes sont égales.

La variable *Longueur* est très importante lors de l'analyse du facteur longueur pour le phénomène d'alternance des compléments en français et a été déjà prouvé comme facteur très important dans les études antérieures (Thuilier, 2012a ; Thuilier, 2012b, Thuilier et al., 2014 ; Thuilier et al., 2021) . D'ailleurs, elle est très utile car représente l'une des manières d'envisager le facteur poids (Faghiri & Thuilier, 2018). (section 2.2)

- Complexité

Cette variable indique le type de SN (complexe ou pas complexe) et SP (complexe ou pas complexe) conformément à l'ordre d'apparition. Je considère un syntagme complexe s'il a une subordonnée ou au moins un SP à l'intérieur. Les valeurs possibles de la variable *Complexité* sont :

- *complexe-pas_complexe* : S'il y a un ordre de type SN-SP dans la phrase et le référent du SN est complexe et le référent du SP n'est pas complexe;
- *pas_complexe-complexe* : S'il y a un ordre de type SN-SP dans la phrase et le référent du SN n'est pas complexe et le référent du SP est complexe;
- *complexe-complexe* : Les référents des deux syntagmes sont complexes;
- pas_*complexe*-pas_*complexe* : Aucun des référents des deux syntagmes n'est complexe.

La variable *Complexité* est importante dans l'analyse de l'alternance des compléments car elle représente une autre manière d'envisager le poids (Faghiri & Thuilier, 2018).

- Définitude

Cette variable indique le type de SN (défini ou indéfini) et SP (défini ou indéfini) conformément à l'ordre d'apparition. Je considère un syntagme comme défini si sa tête est introduite par un article défini et indéfini si sa tête a un article indéfini ou si elle n'a pas d'article. Les valeurs possibles de la variable *Définitude* sont :

- *défini-indéfini* : S'il y a un ordre de type SN-SP dans la phrase, le référent du SN est *défini* et le référent du SP est *indéfini*;

- *indéfini-défini* : S'il y a un ordre de type SN-SP dans la phrase, le référent du SN est *indéfini* et le référent du SP est défini;
- *défini-défini* : Les référents des deux syntagmes sont définis;
- *indéfini-indéfini* : Aucun des référents des deux syntagmes n'est indéfini.

J'ai choisi d'annoter cette variable, car ce facteur peut contribuer à l'étude d'un autre facteur, notamment l'*accessibilité discursive* dans le cas du SN (Faghiri & Thuilier, 2018).

- Pronominalité

Cette variable indique le type de la tête du SN (pronom ou non-pronom) et SP (pronom ou non-pronom) conformément à l'ordre d'apparition. Les valeurs possibles de la variable *Pronominalité* sont :

- *pronom-nonpronom* : S'il y a un ordre de type SN-SP dans la phrase et la tête du SN est un pronom et le dépendant de la tête du SP n'est pas un pronom;
- *nonpronom-pronom* : S'il y a un ordre de type SN-SP dans la phrase, la tête du SN n'est pas un pronom et le dépendant de la tête du SP est un pronom;
- *pronom-pronom* : La tête du SN et le dépendant de la tête du SP sont des pronoms;
- nonpronom-*nonpronom* : Aucune des deux têtes n'est un pronom.

J'ai choisi d'inclure ce facteur car mon étude se concentre sur ce facteur sous plusieurs angles, d'un côté dans le cadre des réalisations des arguments du verbe *donner* et d'un autre côté dans le cadre de l'alternance d'ordre des compléments postverbaux en français oral spontané. En plus, la pronominalité a fait l'objet de plusieurs études sur l'anglais et l'allemand (Bresnan et al., 2007 ; Kempen & Harbusch, 2004 ; Tigău, 2020) et les chercheurs ont pu observer un effet de la pronominalité sur l'ordre des constituants dans la phrase.

- *Ordre*

Cette variable indique l'ordre des syntagmes dans la phrase concernée (SN-SP ou bien SP-SN). Ce facteur a été longuement étudié dans les études antérieures portant sur le phénomène syntaxique d'alternance d'ordre des compléments postverbaux en français. Elle représente le facteur qui m'intéresse le plus, étant donné que tous les facteurs contribuent à l'ordre des compléments.

Les variables suivantes sont des variables obtenues de manière automatique à partir des variables que j'ai annotées semi-automatiquement à l'aide de la plateforme *INCEpTION*.

- Accessibilité

Afin de mieux observer l'interaction entre le facteur *accessibilité discursive* et l'ordre des compléments postverbaux, j'ai créé sous R une nouvelle variable (*Accessibilité*) qui indique le type de SN (accessible, inférable ou inaccessible) et le type du SP (accessible, inférable ou inaccessible) conformément à l'ordre d'apparition. Les valeurs possibles de la variable *Accessibilité* sont :

- *donné-nouveau* : S'il y a un ordre de type SN-SP dans la phrase et la valeur de la variable *SN_accessibilité* est *SN_accessible ou SN_inférable* et la valeur de la variable *SP_accessibilité* est *SP_inaccessible*;
- *nouveau-donné* : S'il y a un ordre de type SN-SP dans la phrase, la valeur de la variable *SN_accessibilité* est *SN_inaccessible* et la variable *SP_accessible* est *SP_accessible* ou *SP_inférable*;
- *donné-donné* : Les valeurs des deux variables sont *SN_accessible* ou *SN_inférable*, la valeur de la variable *Accessibilité* sera *donné-donné*.
- *nouveau-nouveau* : Les valeurs des deux variables sont *SN_inaccessible* et *SP_inaccessible*.

- Animéité

Afin de mieux observer l'interaction entre le facteur *animéité* et l'ordre des compléments postverbaux, j'ai créé sous R une nouvelle variable (*Animéité*) qui indique le type de SN (animé ou inanimé) et SP (inanimé ou inanimé) conformément à l'ordre d'apparition. Les valeurs possibles de la variable *Animéité* sont :

- *animé-inanimé* : S'il y a un ordre de type SN-SP dans la phrase, la valeur de la variable *SN_animéité* est *SN_animé* et la valeur de la variable *SP_animéité* est *SP_inanimé*;
- *inanimé-animé* : S'il y a un ordre de type SN-SP dans la phrase, la valeur de la variable *SN_animéité* est *SN_inanimé* et la variable *SP_animéité* est *SP_animé*;
- *animé-animé* : Les valeurs des deux variables sont *SN_animéité*;
- *inanimé-inanimé* : Les valeurs des deux variables sont *SN_inanimé* et *SP_inanimé*.

La section suivante présente les résultats qui ont été obtenus après l'utilisation des tests statistiques dans le cadre du phénomène syntaxique-ciblé sur les données que j'ai extraites (section 4) et annotées préalablement (section 5).

6 Analyses statistiques de l'ordre des compléments postverbaux en français oral spontané

Cette section présente les résultats obtenus à partir des tests statistiques en ce qui concerne les facteurs qui interagissent avec la variable *Ordre* dans la question de l'alternance des compléments postverbaux en français oral spontané. Tous les résultats ont été obtenus à l'aide du *langage R*[9].

Les facteurs dont les effets ont été testés en interaction avec la variable Ordre sont expliqués dans les lignes qui suivent.

Poids grammatical

Puisque le poids grammatical est envisagé des deux manières : en tant que longueur et en tant que complexité syntaxique, je vais présenter les résultats pour chaque mesure à tour de rôle.

La longueur

En partant de l'hypothèse qu'il n'existe pas de liaison entre les variables *Longueur* et *Ordre*, j'ai effectué le test du Khi-deux après avoir analysé la table des fréquences des valeurs de la variable *Longueur* et la table de contingence croisée entre les deux valeurs qui indique les proportions relatives à chaque combinaison de valeurs.

La répartition des valeurs de la variable *Longueur* au niveau de toutes les données est visible dans le *Tableau 17* :

Valeur de la variable *Longueur*	Nombre de phrases	% dans les données
court-long	64	52.03%
long-court	27	26.02%
autre	32	21.95%

Tableau 17 Répartition des valeurs de la variable Longueur dans les données

Ainsi, la majorité des données se conforment au principe court avant long (52.03%) et 26.02% des données respectent le principe long avant court.

La répartition des valeurs de la variable *Longueur* au niveau de toutes les données en fonction de la variable *Ordre* :

Valeur de la variable *Longueur*	SN-SP	SP-SN
court-long	98.44% (63 phrases)	1.56% (1 phrase)
long-court	74.07% (20 phrases)	25.93% (7 phrases)
autre	100% (32 phrases)	0% (0 phrases)

Tableau 18 La variable Longueur en fonction des valeurs de la variable Ordre

[9] https://www.r-project.org/

L'ordre SN-SP est le plus fréquent dans tous les cas. En revanche, dans le cas de la valeur long-court, il y a plus de phrases dont les arguments du verbe *donner* respectent un ordre de type SP-SN (25.93%) que dans les cas des autres valeurs. Dans les cas *autres*, toutes les phrases présentent un ordre de type SN-SP.

Le test du Khi-deux montre une corrélation entre les deux variables (p-value = 2.098e-05 < 0.05) avec un effet moyen (coefficient V de Cramer = 0.42). La liaison est significative dans le cas des valeurs court-long et long-court par rapport à la variable *Ordre*. Ce résultat va dans le même sens que les résultats des études effectuées jusqu'à présent sur l'ordonnancement des compléments postverbaux en français (Thuilier, 2012a ; Thuilier, 2012b, Thuilier et al., 2014 ; Thuilier et al., 2021).

La complexité syntaxique

En partant toujours du principe qu'il n'y a pas de liaison entre la variable *Complexité* et la variable *Ordre*, j'ai effectué le test du Khi-deux après avoir analysé la table des fréquences des valeurs de la variable *Complexité* et la table de contingence croisée entre les deux valeurs qui indique les proportions relatives à chaque combinaison de valeurs.

La répartition des valeurs de la variable *Complexité* au niveau de toutes les données est visible dans le *Tableau 19* :

Valeur de la variable *Complexité*	Nombre de phrases	% dans les données
pas_complexe-pas_complexe	72	58.54%
pas_complexe-complexe	33	26.83%
complexe-pas_complexe	12	9.76%
complexe-complexe	6	4.88%

Tableau 19 *Répartition des valeurs de la variable Complexité dans les données*

Le cas où les deux syntagmes ne sont pas complexes domine au niveau de toutes les données (58.54%). Dans ce tableau, il peut être aussi observé qu'il y a plus de phrases dans les données qui se conforment au principe pas complexe avant complexe (26.83%) que des phrases qui respectent le principe complexe avant pas complexe (9.76%).

La répartition des valeurs de la variable *Complexité* au niveau de toutes les données en fonction de la variable *Ordre* :

Valeur de la variable *Complexite*	SN-SP	SP-SN
pas_complexe-pas_complexe	98.61% (71 phrases)	1.39% (1 phrase)
pas_complexe-complexe	81.82% (27 phrases)	18.18% (6 phrases)
complexe-pas_complexe	100% (12 phrases)	0% (0 phrases)
complexe-complexe	83.33% (5 phrases)	16.67% (1 phrase)

Tableau 20 *La variable Complexité en fonction des valeurs de la variable Ordre*

L'ordre SN-SP est le plus fréquent dans tous les cas, par contre dans le cas de la valeur *pas_complexe-complexe*, il y a plus de phrases dont les arguments du verbe *donner* respectent un ordre de type SP-SN (18.18%) que dans les cas des autres valeurs. Dans les cas *complexe-pas_complexe*, toutes les phrases présentent un ordre de type SN-SP.

Après avoir fait le test du Khi-deux, les résultats montrent une corrélation entre les variables (p-value = 0.01 < 0.05) avec un effet moyen (coefficient V de Cramer = 0.32). Le résultat montre une liaison significative dans le cas des valeurs *pas_complexe-complexe* et *pas_complexe-pas_complexe* par rapport à la variable *Ordre*. Ce résultat va dans le même sens que le résultat de Faghiri & Thuilier (2018) qui suppose un effet de complexité syntaxique sur l'ordre des compléments postverbeaux en français oral spontané avec une préférence de mettre le syntagme le plus complexe (lourd) vers la fin de la phrase.

Pour résumer, le poids grammatical exerce une influence sur l'ordre des compléments postverbaux en français oral spontané et suppose une préférence pour court avant long et léger avant lourd, ce qui va dans le sens de la littérature dans le cadre de ce phénomène syntaxique. Cependant, le coefficient de Cramer indique un effet un peu plus fort dans le cas de la longueur que dans le cas de la complexité syntaxique (0.42 vs. 0.32), ce qui ne coïncide pas avec les résultats de Faghiri & Thuilier (2018). Puisque mon étude contient un nombre assez réduit de données, toutes ces observations devraient être confirmées dans le cadre d'une autre étude qui implique un nombre plus grand de données.

Définitude

En partant toujours du principe qu'il n'y a pas de liaison entre la variable *Définitude* et la variable *Ordre*, j'ai effectué le test du Khi-deux après avoir analysé la table des fréquences des valeurs de la variable *Définitude* et la table de contingence croisée entre les deux valeurs qui indique les proportions relatives à chaque combinaison de valeurs.

La répartition des valeurs de la variable *Définitude* au niveau de toutes les données est visible dans le *Tableau 21* :

Valeur de la variable *Définitude*	Nombre de phrases	% dans les données
indéfini-défini	65	52.85%
défini-défini	34	27.64%
indéfini-indéfini	19	15.45%
défini-indéfini	5	4.07%

Tableau 21 *Répartition des valeurs de la variable Définitude dans les données*

Ainsi, le cas où les phrases respectent le principe indéfini-défini domine au niveau de toutes les données (52.85%) et les phrases qui respectent le principe défini avant indéfini sont les moins fréquentes.

La répartition des valeurs de la variable *Définitude* au niveau de toutes les données en fonction de la variable *Ordre* :

Valeur de la variable *Définitude*	SN-SP	SP-SN
défini-défini	88.24% (30 phrases)	11.76% (4 phrases)
défini-indéfini	40% (2 phrases)	60% (3 phrases)
indéfini-défini	100% (65 phrases)	0% (0 phrases)
indéfini-indéfini	94.74% (18 phrases)	5.26% (1 phrase)

Tableau 22 *La variable Définitude en fonction des valeurs de la variable Ordre*

En général, l'ordre SN-SP est préféré sauf pour les cas où il y a un SP défini et un SN indéfini (60% vs. 40%). Dans les cas où le SN est indéfini et le SP est défini, toutes les phrases respectent l'ordre SN-SP.

Après avoir fait le test du Khi-deux, les résultats montrent une corrélation entre les variables (p-value = 1.637e-06 < 0.05) avec un effet moyen (coefficient V de Cramer = 0.49) dans le cas des valeurs *défini-indéfini* et *indéfini-défini* par rapport à la variable *Ordre*.

Ces résultats montrent un éventuel effet d'anti-définitude sur l'ordre des compléments postverbaux en français oral spontané avec une préférence de mettre un SN indéfini avant un SP défini, ce qui serait à l'encontre des résultats pour l'anglais et l'allemand qui respectent le principe défini avant indéfini (Bresnan et al., 2007, Kempen & Harbusch, 2004). D'ailleurs, ce résultat ne va pas dans le sens de l'hypothèse de Berrendonner (1987) sur le français qui suppose l'existence du même effet constaté sur l'ordre pour les langues mentionnées antérieurement. Puisqu'il s'agit d'un nombre très réduit de données et assez déséquilibré entre les deux ordres possibles, cette hypothèse pourrait être testé sur un plus grand nombre de données dans le cadre d'une autre étude.

Pronominalité

En partant toujours du principe qu'il n'y a pas de liaison entre la variable *Pronominalité* et la variable *Ordre*, j'ai effectué le test du Khi-deux après avoir analysé la table des fréquences des valeurs de la

variable *Pronominalité* et la table de contingence croisée entre les deux valeurs qui indique les proportions relatives à chaque combinaison de valeurs.

La répartition des valeurs de la variable *Pronominalité* au niveau de toutes les données est visible dans le *Tableau 23* :

Valeur de la variable *Pronominalité*	Nombre de phrases	% dans les données
nonpronom-nonpronom	111	90.24%
nonpronom-pronom	6	4.88%
pronom-nonpronom	6	4.87%
pronom-pronom	0	0%

Tableau 23 *Répartition des valeurs de la variable Pronominalité dans les données*

Ainsi, le cas où il n'y a que des non-pronoms domine au niveau de toutes les données (90.24%). Dans ce tableau, il peut être aussi observé que les phrases ou un syntagme contient un référent qui n'est pas pronom suivi par un syntagme dont le référent est un pronom sont aussi fréquentes que les phrases qui présentent l'ordre inverse.

La répartition des valeurs de la variable *Pronominalité* au niveau de toutes les données en fonction de la variable *Ordre* :

Valeur de la variable *Pronominalité*	SN-SP	SP-SN
nonpronom-nonpronom	92.79% (103 phrases)	7.21% (8 phrases)
nonpronom-pronom	100% (6 phrases)	0% (0 phrases)
pronom-nonpronom	100% (6 phrases)	0% (0 phrases)
pronom-pronom	0% (0 phrases)	0% (0 phrases)

Tableau 24 La variable Pronominalité en fonction des valeurs de la variable Ordre

En général, l'ordre SN-SP est préféré. Ainsi, même s'il s'agit des phrases qui respectent le principe non-pronom avant pronom ou des phrases qui se conforment au principe pronom avant non-pronom, l'ordre SN-SP est toujours préféré.

Après avoir fait le test, les résultats ne montrent pas une corrélation entre les deux variables (p-value = 0.63 > 0.05). Par conséquent, la pronominalité ne semble pas avoir une influence sur l'ordre des compléments postverbaux en français oral spontané, ce qui va dans le sens des résultats de Thuilier (2012a, b). Puisque mon étude contient un nombre assez réduit de données, toutes ces observations devraient être confirmées dans le cadre d'une autre étude qui implique un nombre plus grand de données.

Accessibilité discursive

En partant du principe qu'il n'y a pas de liaison entre la variable *Accessibilité* et la variable *Ordre*, j'ai effectué le test du Khi-deux après avoir analysé la table des fréquences des valeurs de la variable *Accessibilité* et la table de contingence croisée entre les deux valeurs qui indique les proportions relatives à chaque combinaison de valeurs.

La répartition des valeurs de la variable *Accessibilité* au niveau de toutes les données est visible dans le *Tableau 25* :

Valeur de la variable *Accessibilité*	Nombre de phrases	% dans les données
nouveau-donné	67	54.47%
donné-nouveau	55	44.72%
nouveau-nouveau	1	0.81%
donné-donné	0	0%

Tableau 25 *Répartition des valeurs de la variable Accessibilité dans les données*

Les phrases qui se conforment au principe nouveau avant donné représentent la majorité des données (54.47%). Dans ce tableau, il peut être aussi observé que le fait de mettre un référent donné avant un référent nouveau est aussi très fréquent (44.72%).

La répartition des valeurs de la variable *Accessibilité* au niveau de toutes les données en fonction de la variable *Ordre* :

Valeur de la variable *Accessibilité*	SN-SP	SP-SN
nouveau-donné	89.09% (65 phrases)	2.99% (2 phrases)
donné-nouveau	97.01% (49 phrases)	10.91% (6 phrases)
nouveau-nouveau	100% (1 phrase)	0% (0 phrases)
donné-donné	0% (0 phrases)	0% (0 phrases)

Tableau 26 *La variable Accessibilité en fonction des valeurs de la variable Ordre*

L'ordre SN-SP est le plus fréquent dans tous les cas, en revanche dans le cas de la valeur nouveau-donné, il y a plus de chances d'avoir un ordre SP-SN (10.91%) que dans les cas des autres valeurs.

Après avoir fait le test du Khi-deux, les résultats ne montrent pas une corrélation entre les variables (p-value = 0.2 > 0.05). Par conséquent, il est possible que l'accessibilité discursive n'influence pas l'ordre des compléments postverbaux en français oral spontané. Ce résultat ne coïncide pas avec les résultats de Faghiri & Thuilier (2018). Le résultat n'identifie pas l'existence d'une liaison significative entre les deux variables, ce qui va à l'encontre de mon hypothèse de départ selon laquelle l'effet de l'accessibilité discursive serait plus visible à l'oral. Puisque mon étude contient un nombre assez réduit

de données, toutes ces observations devraient être confirmées dans le cadre d'une autre étude qui implique un nombre plus grand de données.

Animéité

En partant toujours du principe qu'il n'y a pas de liaison entre la variable *Animéité* et la variable *Ordre*, j'ai effectué le test du Khi-deux après avoir analysé la table des fréquences des valeurs de la variable *Animéité* et la table de contingence croisée entre les deux valeurs qui indique les proportions relatives à chaque combinaison de valeurs.

La répartition des valeurs de la variable *Animéité* au niveau de toutes les données est visible dans le *Tableau 27* :

Valeur de la variable *Animéité*	Nombre de phrases	% dans les données
inanimé-animé	83	67.48%
inanimé-inanimé	33	26.83%
animé-inanimé	7	5.69%
animé-animé	0	0%

Tableau 27 *Répartition des valeurs de la variable Animéité dans les données*

Le cas où un syntagme dont le référent est inanimé précède un syntagme dont le référent est animé domine au niveau de toutes les données (67.48%). Dans ce tableau, il peut être aussi observé qu'il y a aussi des phrases dans les données qui se conforment au principe anime avant inanimé mais elles sont beaucoup moins fréquentes (5.69%).

La répartition des valeurs de la variable *Animéité* au niveau de toutes les données en fonction de la variable *Ordre* :

Valeur de la variable *Animéité*	SN-SP	SP-SN
animé-inanimé	0% (0 phrases)	100% (7 phrases)
inanimé-animé	100% (83 phrases)	0% (0 phrases)
inanimé-inanimé	96.97%(32 phrases)	3.03%(1 phrase)
animé-animé	0% (phrases)	0% (phrases)

Tableau 28 *La variable Animéité en fonction des valeurs de la variable Ordre*

L'ordre SN-SP est le plus fréquent dans presque tous les cas sauf pour les cas de la valeur *animé-inanimé* où il y a exclusivement l'ordre SP-SN. Les cas où un syntagme dont le référent est inanimé précède un syntagme dont le référent animé se retrouvent exclusivement dans l'ordre SN-SP.

Après avoir fait le test du Khi-deux, les résultats montrent une corrélation entre les variables (p-value = 2.2e-16 < 0.05) avec un effet très fort (coefficient V de Cramer = 0.93). Le résultat montre donc une forte liaison dans le cas des valeurs *inanimé-animé* et *animé-inanimé* par rapport à la variable *Ordre*. Ce résultat va dans le même sens que mon hypothèse de départ qui suppose un effet

d'anti-animéité sur l'ordre des compléments postverbaux en français oral spontané avec une préférence de mettre un OD inanimé avant un OI animé.

Dans les cas de tous les tests, il est possible que les résultats ne soient pas correctement estimés à cause des effectifs très bas et pas assez équilibrés entre les deux ordres des syntagmes concernés (115 phrases qui ont un ordre de type SN-SP contre 8 phrases qui présentent un ordre de type SP-SN), ce qui indique une préférence dans le cadre du phénomène d'alternance d'ordre des compléments postverbaux en français oral spontané pour un ordre de type SN-SP plutôt qu'un ordre de type SP-SN dans le cas du verbe *donner*. Il serait intéressant de retester les mêmes effets sur un nombre de données plus élevé et plus équilibrée du point de vue de l'ordre des syntagmes.

Les cas de phrases contenant le verbe donner en tant que verbe support (ex. *donner confiance à qqn.*, *donner satisfaction à qqn.*) représentent presque la majorité des données (47.15%) et elles supposent dans la plupart des cas un syntagme court/léger dont le référent peut être inanimé et défini avant un syntagme long/lourd dont le référent peut être animé et défini. C'est ainsi qu'il est possible que cette variable ait une influence sur les résultats obtenus.

La réalisation d'un modèle de régression linéaire à effets mixtes serait souhaitable afin de voir s'il y a des interactions supplémentaires entre les variables ou des variables dont les effets se neutralisent et de généraliser les résultats à partir de l'échantillon étudié.

La section suivante propose une discussion sur tous les résultats des analyses effectuées et illustre les limites de cette étude intervenant dans chaque étape de la chaîne de traitement mise en place.

7 Discussion

Cette étude fournit des éléments en faveur de l'éventuelle existence d'une série de contraintes préférentielles et leurs effets sur l'ordre des compléments postverbaux en français oral spontané dans le cas du verbe *donner*. Puisque la pronominalité n'a pas pu être observé statistiquement dans les études antérieures sur le français contrairement à l'anglais et à l'allemand à cause de la présence réduite des pronoms dans la zone postverbale, l'un des objectifs de ce mémoire a été de décrire toutes les réalisations des arguments du verbe *donner* qui incluent aussi des pronoms clitiques.

La démarche expérimentale que j'ai mise en place dans ce mémoire a compris quatre étapes principales. Cette démarche expérimentale a apporté de nouveaux éléments en ce qui concerne la préparation de l'analyse et l'analyse en elle-même des données provenant de la modalité orale. Les quatre étapes ont permis de dépasser certaines limites apparues à cause de l'annotation automatique du corpus contenant des erreurs. Cependant, il reste des limites malgré toutes les manipulations entreprises. Tous ces aspects sont décrits à tour de rôle pour chaque grande étape dans les lignes qui suivent.

La sélection des données

Cette première étape a été très importante afin d'effectuer mes analyses sur des données provenant exclusivement des locuteurs natifs du français et contenant de l'oral spontané. Ainsi, cette étude se différencie par rapport aux autres études effectuées sur le français qui supposent des données appartenant exclusivement à la modalité écrite ou contiennent les deux modalités. Dans le cas des études antérieures, la modalité orale ne fait pas l'objet des discours spontanés, mais plutôt des discours plus ou moins planifiés, comme les émissions radio. Mon étude se base sur neuf sous-corpus inclus dans la partie orale du corpus *CEFC* (*CRFP*, *CFPP*, *CLAPI*, *VALIBEL*, *TCOF*, *Réunions-de-travail*, *TUFS*, *CFPB* et *OFROM*) qui contiennent de diverses situations d'oral spontané, comme des enregistrements avec des locuteurs natifs dans des situations naturelles et des contextes variés.

La constitution du jeu de données

Un point important a été représenté par la méthode d'extraction qui a demandé la création des scripts informatiques beaucoup plus efficaces au niveau du temps par rapport à l'extraction directe depuis la plateforme *Orféo*. Ce mémoire a également montré comment mettre en place une chaîne de traitement qui permette la constitution et l'annotation d'un jeu de données de haute qualité nécessaire pour l'analyse de l'alternance d'ordre des compléments postverbaux en français oral spontané. La chaîne de traitement a supposé une re-annotation des phrases contenant le lemme *donner* avec le parser *Stanza* afin de réduire le nombre d'erreurs d'annotations qui aurait été difficile à contourner même avec un script à cause du bruit engendré. Cette procédure a permis la réduction des erreurs d'annotations qui ne sont pas récurrentes (donc qui ne peuvent pas être contournées avec un script, car cela pourrait engendrer beaucoup de bruit) et l'élimination du problème représenté par le déterminant *des* étiqueté en tant que préposition. Cependant, il reste toujours quelques erreurs d'annotation isolées et des erreurs générées par l'existence des disfluences du langage oral qui ont été transcrites et conservées. Par conséquent, les hésitations ou les auto-corrections peuvent conduire à des doublons dans l'identification des structures-cible.

La méthode d'évaluation du Gold standard a permis de mesurer la qualité de l'extracteur et pour l'améliorer au fur et à mesure. Toutefois, il est impossible d'estimer son efficacité précise une fois appliqué au corpus entier car des données peuvent être passées sous silence du fait d'erreurs d'annotation et de leur absence dans le sous-corpus à partir duquel le Gold a été réalisé (le sous-corpus *CRFP*).

Afin d'assurer une haute qualité au niveau de l'extraction des données, une vérification manuelle a été requise après toutes les manipulations automatiques effectuées.

Une fois les données récupérées et vérifiées manuellement, j'ai analysé toutes les réalisations des arguments du verbe donner autant au niveau du sous-corpus CRFP que du corpus CEFC sélectionné. Ce qui a été intéressant a été le fait que les proportions de types de structures sont restées similaires après avoir entendu l'extracteur au niveau du corpus entier sélectionné. Ainsi, pour toutes les deux situations, les phrases qui constituent la majorité des données ne contiennent aucune structure-cible. Une autre similarité est que les phrases contenant des structures de type B (celles qui ont des arguments après et avant le verbe) sont les plus fréquentes et ainsi, les phrases contenant des structures de type B et C (donc les structures qui contiennent des pronoms clitiques) sont plus nombreuses que les phrase de type A (ne contenant pas de clitiques et dont les arguments sont réalisés exclusivement dans la zone postverbale). En effet, le nombre de données pertinentes pour l'alternance d'ordre des compléments postverbaux en français oral spontané est assez réduit dans tous les deux cas (6.5% dans le sous-corpus *CRFP* et 4.1% dans le corpus *CEFC* sélectionné). Par conséquent le nombre très élevé de structures contenant des pronoms clitiques (34.62% dans le sous-corpus CRFP et 27.94% dans le corpus CEFC) ont permis de voir une certaine cohérence et une quantification au niveau des tendances du français de mettre les pronoms plutôt dans la zone préverbale que dans la zone postverbale. Un autre élément intéressant a été l'observation des structures de type B qui contiennent l'un des deux arguments disloqués dans la zone postverbale (21 phrases indépendamment du type d'argument concerné).

L'annotation des données pour l'étude de l'alternance des compléments postverbaux en français oral spontané

L'annotation des données a été réalisée de manière automatique par l'intermédiaire d'un script informatique et semi-automatique par l'utilisation de la plateforme d'annotation *INCEpTION*.

Dans le cas de l'annotation automatique, les difficultés qui sont restées après la re-analyse des données avec Stanza provenaient des erreurs d'annotations mentionnées ci-dessous, fait qui a rendu le processus d'annotation plus difficile, car le plus d'erreurs possibles devraient être contournées pour obtenir une annotation fiable pour les analyses statistiques.

Dans le cas de l'annotation semi-automatique, le processus a été plus efficace qu'une annotation manuelle grâce à l'outil d'apprentissage automatique *Recommenders* existant sur la plateforme *INCEpTION*. Les scores des mesures en ce qui concerne l'efficacité de l'outil différent, car les couches *SN_animéité* ou *SP_animéité* s'apprêtaient mieux à l'identification de la séquence ciblée que les couches créées pour l'accessibilité discursive. Cet aspect a été possible car il y a plus de chances de retrouver le même référent et qu'il soit toujours animé, par exemple, que de retrouver le même référent qui a le même rôle discursif dans une autre phrase.

Une limite restante dans l'annotation semi-automatique réalisée a été la subjectivité de l'annotation. (Par exemple dans le cas de l'accessibilité discursive : l'hésitation entre un référent accessible, inférable ou pas du tout accessible). Cette subjectivité pourrait être réduite dans de futurs travaux en passant par une annotation à plusieurs annotateurs et par la rédaction d'un guide d'annotation où les catégories à annoter et les choix à faire dans les cas difficiles sont exposés, pour essayer de conserver un maximum de cohérence.

Les analyses statistiques sur l'alternance d'ordre des compléments postverbaux en français oral spontané

Dans ce mémoire, j'ai pu observer s'il y a des corrélations entre plusieurs variables (*Longueur*, *Complexité*, *Définitude*, *Pronominalité*, *Accessibilité* et *Animéité*) et la variable Ordre, afin d'identifier des hypothèses sur des éventuelles contraintes préférentielles qui pourraient intervenir dans le phénomène d'alternance d'ordre des compléments postverbaux en français oral spontané. En partant

du principe qu'il n'existe pas de liaison significative entre les variables mentionnées et après avoir effectué le test du Khi-deux, j'ai pu observer des corrélations ou l'absence des corrélations entre les variables étudiées.

En ce qui concerne le poids grammatical, j'ai pu voir une corrélation avec un effet moyen entre la variable *Longueur* et la variable *Ordre*. Les paires significatives de valeurs ont indiqué une préférence de mettre un syntagme court avant un syntagme long et des cas plus fréquents avec ordre de type SP-SN lorsqu'un syntagme long précède un syntagme plus court. Quant à la complexité discursive, j'ai pu observer une corrélation entre la variable *Complexité* et la variable *Ordre* avec une préférence de mettre un syntagme moins complexe (léger) avant un syntagme plus complexe (plus lourd). Contrairement à l'étude de Faghiri & Thuilier (2018), selon laquelle l'effet de complexité serait plus fort que l'effet de longueur, dans mon cas, la corrélation semble un peu plus forte pour l'effet de longueur que pour l'effet de complexité syntaxique.

Quant au caractère défini, j'ai pu observer une corrélation entre la variable *Définitude* et la variable *Ordre* avec la tendance de mettre un syntagme dont le référent est inanimé avant un syntagme dont le référent est animé dans le cas de l'ordre SN-SP. D'ailleurs, quand le SP est défini et le SN est indéfini, il existe plus de phrases suivant un ordre de type SP-SN (60%). Cependant, comme il s'agit de très peu de phrases et la différence entre le nombre de phrases est minimale en ce qui concerne une préférence de type SP défini avant un SN indéfini, cet aspect devrait être testé sur un échantillon plus grand. Toutes ces observations pourraient soutenir l'hypothèse sur l'existence d'un certain effet d'anti-définitude dans le cas de l'oral spontané qui pourrait être vérifiée dans le cadre des autres travaux.

En ce qui concerne la pronominalité, je n'ai pas pu observer une corrélation ou une certaine tendance sur la variable *Ordre* à cause du nombre insuffisant de données.

Contrairement à mon hypothèse de départ, selon laquelle l'effet d'accessibilité discursive serait plus évident à l'oral, je n'ai pas pu constater une corrélation entre les variables *Accessibilité* et *Ordre*. Ce résultat ne va pas dans le sens du résultat de l'étude Faghiri & Thuilier (2018).

Concernant l'animéité, j'ai pu constater une liaison significative avec un effet fort entre les variables *Animéité* et *Ordre*. Les paires significatives de valeurs ont indiqué une préférence de mettre un SN inanimé avant un SP animé dans le cas d'un ordre de type SN-SP et une préférence pour un SP animé avant un SN inanimé pour l'ordre SP-SN. Ces résultats vont dans le sens de l'étude de Thuilier et al. (2021) qui ont montré l'existence d'un effet d'anti-animéité.

Puisque le nombre de phrases contenant le verbe donner en tant que verbe support est assez élevé, il serait intéressant d'observer l'effet de cette variable dans le cadre du phénomène d'alternance des compléments postverbaux en français oral spontané pour une plus grand nombre de données. C'est ainsi qu'il est possible que cette variable ait une influence sur les résultats obtenus. Une solution pour se rendre compte du rôle que cette variable peut jouer serait d'écarter les données contenant le verbe donner de type support et de refaire les tests afin de voir si les mêmes tendances se maintiennent et si l'intensité des effets dans le cas des corrélations change ou bien il n'existe plus de corrélations entre les variables. Dans le cas de mon étude, une telle démarche ne pourrait pas s'appliquer à cause du nombre insuffisant de données qui affectent la significativité des tests statistiques.

Tous ces résultats et hypothèses devraient être testés dans le cadre d'une étude disposant d'un plus grand nombre de données. Il serait préférable de mettre en place un modèle de régression logistique à effets mixtes afin de voir le poids exact de chaque facteur intervenant, de constater si les effets de certaines variables se neutralisent en interaction ou une variable renforce l'effet d'une autre variable sur le facteur Ordre et de généraliser au-delà de l'échantillon étudié.

8 Conclusions et perspectives

La mise en place d'une chaîne de traitement et l'utilisation des méthodes automatiques et semi-automatiques suivies par une vérification manuelle peuvent assurer une haute qualité et efficacité dans la constitution et l'annotation d'un jeu de données, étapes nécessaires dans l'analyse d'un phénomène syntaxique étudie, comme l'alternance d'ordre des compléments postverbaux en français oral spontané.

Selon les résultats des tests statistiques effectués dans le cadre de cette étude, l'oral spontané semble apporter des éléments nouveaux en ce qui concerne l'ordre des compléments postverbaux en français. Ainsi, les contraintes préférentielles et leurs effets sur l'ordre des arguments du verbe peuvent différer par rapport à la modalité écrite. Il est important de mentionner que toutes les analyses ont été menées exclusivement sur le verbe *donner* (qui semble encourager un ordre de type SN-SP) et dans le cas des autres verbes appartenant aux autres classes sémantiques, les résultats pourraient changer.

L'insuffisance des données ne permet pas malheureusement la possibilité de tirer des conclusions sur l'existence des effets en ce qui concerne le phénomène syntaxique étudié. Un autre problème engendré par le nombre réduit des données a été le fait de ne pas pouvoir observer certains facteurs, comme la pronominalité. Il serait intéressant d'étudier dans de futurs travaux la liaison entre la position d'un argument dont le référent est un pronom et l'ordre des compléments du verbe. Il est possible que d'autres facteurs encouragent la présence des pronoms dans la zone préverbale par l'intermédiaire des pronoms clitiques. Cela pourrait être possible s'il y a nombre suffisant des données qui contiennent des pronoms dans la zone postverbale. En plus, la mise en place d'un modèle de régression logistique à effets mixtes (comme dans le cas des études de Thuilier, 2012a, b et de Bresnan et al., 2007) permettrait d'observer le poids de chaque facteur impliqué, toutes les interactions qui pourraient avoir lieu entre les variables et de généraliser au-delà de l'échantillon étudié. Dans une étude future, l'étude sur corpus pourrait être accompagnée par une tâche des jugements d'acceptabilité réalisée auprès des locuteurs natifs en contrôlant la position des pronoms, afin d'observer le facteur *pronominalité*.

Une autre piste intéressante serait d'étendre l'analyse sur plusieurs lemmes verbaux appartenant à la classe des verbes de transfert (similaires au verbe *donner*) ou même à toutes les classes sémantiques génériques prises déjà en considération par Thuilier (2012a, b), fait qui permettrait d'introduire encore trois variables, notamment le lemme verbal, la classe sémantique et le rôle sémantique des arguments.

Une autre possibilité pour de futurs travaux serait de faire une comparaison entre les facteurs et leurs effets identifiés dans le cadre de la modalité orale et de la modalité écrite afin de se rendre compte encore mieux de l'influence que l'oral puisse exercer sur l'ordre des compléments du verbe. Cette analyse comparative pourrait être menée sur le corpus *CEFC* étant donné le fait qu'il contient autant de la modalité orale que de la modalité écrite.

Bibliographie

Abeillé A. & Godard D. (2004). De la légèreté en syntaxe. *Bulletin de la Société de Linguistique de Paris*, XCIX(1), 69–106.

Abeillé A. & Godard D. (2006). La légèreté en français comme déficience de mobilité. *Lingvisticae Investigationes*, 29(1), 11–24.

Abeillé, A., Godard, D., Delaveau, A. & Gautier, A. (2021). *La Grande Grammaire du français*. Paris : Actes Sud.

Adam, C., Fabre, C. et al. (2013). Évaluer et améliorer une ressource distributionnelle : Protocole d'annotation de liens sémantiques en contexte. *Revue TAL*, *54* (1) , 71-97.

Bîlbîie, Faghiri & Thuilier. (2021), Syntaxe quantitative et expérimentale : objets et méthodes, *Langages*, 223, 7-24.

Benzitoun, C., Debaisieux, J.-M., & Deulofeu, H.-J. (2016). Le projet ORFÉO : Un corpus d'étude pour le français contemporain. *Corpus*, *15*.

Berrendonner A. 1987. L'ordre des mots et ses fonctions. *Travaux de linguistique*, 14/15, 9–19.

Bérard, L. (2020). La partie orale du Corpus d'Étude pour le Français Contemporain (CEFC). *Langages (Paris), 219*(3), 25-37.

Blinkenberg, A. (1928). L'ordre des mots en français moderne, première partie, Copenhague : Høst & Søn.

Bresnan, J., Cueni, A., Nikitina, T. & Baayen, H. (2007). Predicting the dative alternation. In G. Boume, I. Kraemer & J. Zwarts, Eds., *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science.

Bresnan J., Dingare S. & Manning C. D. (2001). Soft constraints mirror hard constraints : Voice and person in English and Lummi. In M. Butt & T. H. King, Eds., *Proceedings of the LFG01 Conference*, Hong Kong.

Bresnan J. & Ford M. (2010). Predicting syntax : Processing dative constructions in American and Australian varieties of English. *Language*, 86(1), 186–213.

Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge: MIT Press.

Chomsky, N (1975). The logical structure of Linguistic Theory. Cambridge: MIT Press.

Da Cunha, Y. (2021). L'alternance actif/passif en français écrit et oral : études quantitatives et formelles. [Mémoire du master, Université de Paris]

Dobrescu, A. M., Thuilier, J, Fabre, C. & Bîlbîie, G. (2024). L'extraction des données du corpus CEFC pour l'analyse de l'alternance d'ordre des compléments en français oral spontané. *Langue et Culture*, *3* [à paraître, Université de Bucarest].

Dubois, J. & Dubois-Charlier, F. (1997). *Les verbes du français*. Paris : Larousse-Bordas.

Enkvist, N.E. (1985). A Parametric View of Word Order. In E. Sözer, Ed. *Text Connexity, Text Coherence*, 320-336. Hamburg: Helmut Buske.

Firbas, J. (1972). On the Interplay of Prosodic and Non-Prosodic Means of Functional Sentence Perspective. In U. Fried, Ed. *The Prague School of Linguistics and Language Teaching*, 77-94. London: Oxford University Press.

Faghiri, P. & Thuilier, J. (2018). Ordre des compléments postverbaux en français : Poids et accessibilité discursive. In F. Neveu, B. Harmegnies, L. Hriba & S. Prévost (eds.), *SHS web conference : 6ème congrès mondial de linguistique française*.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5, 1–26.

Jelinek, E. & Demers, R. A. (1983). The agent hierarchy and voice in some Coast Salish languages. *International Journal of American Linguistics*, 49(2), 167–185.

Jelinek E. & Demers R. A. (1994). Predicates and pronominal arguments in Straits Salish. *Language*, 70(4), 697–736.

Kempen, G. & Harbusch, K. (2004). A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In T. Pechmann, et C. Habel (eds.), *Multidisciplinary approaches to language production*, 173-181. Berlin : Mouton de Gruyter.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R. & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico, USA.

Nasr, A., Béchet, F., Rey, J.-F., Favre, B., & Roux, J. (2011). *MACAON An NLP Tool Suite for Processing Word Lattices*. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, 86–91, Portland, Oregon. Association for Computational Linguistics.

Péry-Woodley, M. (2000). Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. Linguistique. Chapitre 2, 18-24. [Habilitation à diriger des recherches, Université Toulouse le Mirail - Toulouse II]

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole, Ed., Radical *Pragmatics*, 223–256. New York : Academic Press.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Schmitt, C. (1987). A propos de l'impact de la sémantique sur la séquence des compléments d'objet en français moderne. *Travaux de linguistique et de littérature* 25(1), 283-298.

Thuilier, J. (2012a). Lemme verbal et classe sémantique dans l'ordonnancement des compléments postverbaux. In *Actes du congrès mondial de linguistique française 2012* (cmlf 2012).

Thuilier, J. (2012b). Contraintes préférentielles et ordre des mots en français : Université Paris 7 [Thèse de doctorat].

Thuilier, J. et al. (2014). Ordering preferences for postverbal complements in French. *In Henry* Tyne, Virginie André, Christophe Benzitoun, Alex Boulton, Yan Greub (dir.), *French through Corpora - Ecological and Data-Driven Perspectives in French Language Studies* . Newcastle upon Tyne, UK : Cambridge Scholars Publishing.

Thuilier, J., Grant, M, Crabbé, B. & Abeillé, A. (2021). Word order in French: The role of animacy. *Glossa: a journal of general linguistics*, 6(1), 55.

Tigău, A. (2020). Experimental Insights of Romanian Ditransitives, Studies in Generative Grammar, Vol. 141, 165-276.

Wang I., Pelletier A., Antoine J.-Y., Halftermeyer, A. (2020). ODIL_Syntax: a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees. *Proceedings of the twelfth Language Resources and Evaluation Conference (LREC'2020)*. Marseille, France.

Zaharlick A. (1982). Tanoan studies : Passive sentences in Picuris. *The Ohio State University Working Papers in Linguistics*, 26, 34–48.

Déclaration sur l'honneur de non-plagiat

(à joindre au mémoire à la fin du document)

Je soussigné.e,

Nom, Prénom : Dobrescu Anca-Mihaela
Régulièrement inscrit.e à l'Université de Toulouse II Jean Jaurès
N° étudiant : 22213472

Année universitaire : 2022-2023

certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à : Toulouse

Le : 20 août 2023

Signature :