Complexité syntaxique et flux de dépendance : études quantitatives dans les treebanks universal dependencies

Chunxiao Yan

Université Paris Nanterre

200 av. de la République
92001 Nanterre Cedex
www.parisnanterre.fr

École doctoral 139
Connaissances, Langage, Modélisation

Membre de l'université Paris Lumières

Chunxiao YAN

Complexité Syntaxique et Flux de Dépendance

Études quantitatives dans les treebanks Universal Dependencies

Thèse présentée et soutenue publiquement le 01/12/2021
en vue de l'obtention du doctorat de Sciences du langage de l'Université Paris Nanterre
sous la direction de M. Sylvain Kahane (Université Paris Nanterre)

Jury :

Rapporteur·e :	Pr François Lareau	Professeur, Université de Montréal
Rapporteur·e :	Pr Philippe Blache	Professeur, Université d'Aix-Marseille
Membre du jury :	Dr Marie Candito	Université Paris-Diderot
Membre du jury :	Pr Marie-Catherine de Marneffe	Université D'État de l'Ohio
Membre du jury :	Pr Kim Gerdes	Professeur, Université Paris-Saclay
Directeur :	Pr Sylvain Kahane	Professeur, Université Paris Nanterre

Résumé

Nous nous intéressons à la complexité syntaxique et aux contraintes liées à la mémoire de travail chez l'humain. La mémoire de travail concerne non seulement la capacité de retenir des informations, mais aussi la capacité de les manipuler temporairement. Elle a été montrée limitée à 7±2 éléments (Miller, 1956) et est aujourd'hui actualisée autour de 4 selon Cowan (2001). La limitation de la mémoire de travail peut rendre le traitement de certaines structures de phrase difficile voire impossible (Holmes & O'Regan, 1981 ; King & Just, 1991 ; Gibson, 1998). Dans cette thèse, nous nous penchons sur trois pistes d'étude : étudier et mesurer la complexité syntaxique sous différentes hypothèses cognitives ; savoir s'il existe des limites à la complexité syntaxique dans les langues naturelles ; comprendre les phénomènes impliqués par les contraintes sur la complexité syntaxique.

De ce fait, nous mesurons la complexité syntaxique en utilisant des métriques basées sur le flux de dépendance (Kahane, 2001) dans le corpus (les treebanks *Universal Dependencies*). Ces métriques incluent non seulement des métriques devenues classiques comme la longueur de dépendance (Liu, 2008 ; Futrell et al, 2015), des métriques proposées dans des travaux plus récents (Kahane et al., 2017), mais aussi de nouvelles métriques également basées sur le flux de dépendance ; en se basant sur les résultats donnés par ces différentes métriques sur les plus de 100 langues appartenant à la collection des treebanks *Universal Dependencies* (Nivre et al., 2016 ; Nivre et al., 2017), nous pouvons déterminer celles qui sont les plus appropriées pour étudier la complexité syntaxique. Nous montrons qu'il existe pour certaines métriques du flux des contraintes universelles, dont nous postulons qu'elles sont liées à la mémoire de travail. Enfin, nous essayons également d'expliquer certains des phénomènes linguistiques observés dans nos données qui impliquent la complexité syntaxique.

Mots clés : flux de dépendance, syntaxe de dépendance, complexité syntaxique, métrique, mémoire de travail, *Universal Dependencies Treebanks*

Abstract

We are interested in syntactic complexity and its constraints on human working memory. Working memory requires not only the ability to retain information, but also the ability to manipulate it temporarily. It was shown to be limited to 7±2 elements (Miller, 1956) and is now updated to around 4 according to Cowan (2001). The limitation of working memory can make the processing of certain sentence structures difficult or impossible (Holmes & O'Regan, 1981; King & Just, 1991; Gibson, 1998). In this thesis, we focus on three areas of study: studying and measuring syntactic complexity under different cognitive assumptions; finding out if there are limits to syntactic complexity in natural languages; and understanding the phenomena involved in constraints on syntactic complexity.

Considering this, we measure syntactic complexity using metrics based on dependency flux (Kahane, 2001) in the corpus (the Universal Dependencies treebanks). These metrics include not only metrics that have become classic such as dependency length (Liu, 2008; Futrell et al., 2015), metrics proposed in more recent work (Kahane et al., 2017), but also new metrics also based on dependency flux; based on the results given by these different metrics on the more than 100 languages belonging to the collection of treebanks Universal Dependencies (Nivre et al., 2016; Nivre et al., 2017), we can determine which ones are the most appropriate to study syntactic complexity. We show that for some flux metrics there are universal constraints, which we postulate are related to working memory. Finally, we also try to explain some of the linguistic phenomena observed in our data that involve syntactic complexity.

Keywords: dependency flux, dependency syntax, syntactic complexity, metric, working memory, *Universal Dependencies Treebanks*

Remerciement

Tout d'abord, je tiens à remercier sincèrement mon directeur de thèse Sylvain Kahane pour sa patience, ses conseils importants et précieux sur mon travail de thèse, ainsi que la relecture et les commentaires des chapitres de ma thèse. Je voudrais également remercier Guillaume Desagulier, Kim Gerdes et Benoît Crabbé d'avoir fait partie de mon comité de suivi de thèse ainsi que tous les collègues du laboratoire Modyco pour nos échanges académiques.

Je tiens à remercier *Chinese Scholarship Council* (CSC) pour le financement de ma thèse.

Je remercie beaucoup François Lareau et Phillipe Blache pour avoir accepté d'être mes rapporteurs de thèse, ainsi que Marie Candito, Marie-Catherine de Marneffe, et Kim Gerdes pour m'avoir fait l'honneur d'être mes examinateurs.

Au cours de la rédaction, j'ai rencontré des difficultés avec le français, je tiens à exprimer ma gratitude à mes collègues doctorantes Esther Gettler et Laura Noreskal pour leur aide généreuse dans la relecture de ma thèse. Enfin, je tiens à remercier mes amis Octave Buffi, Sihang Ma et Xinying Chen, ainsi que tous les membres de ma famille pour leurs encouragements et leur soutien.

Table des matières

0 Introduction Générale

0.1 Problématique

La difficulté de compréhension d'une phrase peut être liée à divers facteurs, tels que des aspects pragmatiques liés au contexte de la conversation, comme les connaissances partagées, ainsi que des aspects plus linguistiques, comme la familiarité avec le vocabulaire utilisé ou, pour ce qui nous intéresse directement, la complexité de la forme structurelle de la phrase. Différents domaines de recherche s'intéressent à la difficulté de compréhension d'une phrase : dans les études d'acquisition, des phrases de différents niveaux de difficulté peuvent être des ressources pour évaluer le degré d'acquisition d'une langue ; de telles phrases peuvent également permettre d'évaluer les compétences linguistiques dans des situations de troubles du langage causés par certaines maladies ; enfin, pour l'apprentissage d'une langue étrangère, elle peut servir de référence aux enseignants pour préparer des supports de cours ainsi que pour évaluer leurs résultats d'apprentissage.

Parmi les nombreux aspects qui influencent la difficulté de compréhension des phrases, notre thèse s'intéresse à l'effet de la **syntaxe**[1]. En d'autres termes, le cœur de nos études porte sur la **complexité syntaxique**, qui se réfère à l'évaluation de la difficulté d'analyser des constructions syntaxiques afin de parvenir à la compréhension pour un locuteur.

0.1.1 Hypothèses

Dans le cadre du traitement des phrases, la complexité syntaxique est associée à certaines hypothèses cognitives. Nous soutenons que la difficulté du traitement de différentes

[1] Nous donnons ici la définition de la syntaxe de Kahane et Gerdes (2021) : *« La syntaxe est l'étude de la combinatoire des unités lexicales et grammaticales et tout particulièrement des combinaisons libres obéissant à des règles générales.* »

combinaisons (des unités syntaxiques) est liée à l'efficacité de leur traitement dans la mémoire de travail[2]. La mémoire de travail nous permet de retenir les mots prononcés par notre interlocuteur et de les manipuler inconsciemment. Pour atteindre la compréhension finale, nous ne semblons pas avoir besoin de retenir chaque mot de la phrase progressivement tout au long du traitement. La mémoire de travail ne peut conserver que les informations nécessaires en les traitant pendant l'écoute. Ces informations sont essentielles pour la construction de la structure éventuelle de la phrase. Nous pensons que le volume d'informations à mémoriser pour parvenir à la compréhension finale varie en fonction de la structure des phrases.

Selon les recherches psychologiques sur la mémoire de travail, il existe des limites sur celle-ci : d'après Miller (1956), le nombre d'éléments que l'humain peut garder temporairement est limité entre 5 et 9 ; ce nombre a été actualisée par Cowan (2001) à 4. Ainsi, non seulement des structures de phrases différentes entraînent des degrés de complexité syntaxique différents, mais il y aurait une limite à la complexité syntaxique des phrases. Les structures complexes qui dépassent cette limite ne peuvent pas être comprises par le locuteur. Nous illustrons cela à l'aide des deux exemples suivants.

Premièrement, plusieurs études (Holmes & O'Regan, 1981 ; King & Just, 1991 ; Gibson, 1998) ont montré que les phrases en anglais qui ont une proposition relative avec extraction du sujet, comme (1a), sont plus faciles à traiter que celles avec extraction de l'objet, comme (1b).

(1) (a). Extraction du sujet: *The reporter that attacked the senator admitted the error.*
(b). Extraction de l'objet: *The reporter that the senator attacked admitted the error.*
King et Just (1991)

[2] Du point de vue cognitif, la mémoire de travail est un système cognitif à capacité limitée qui peut retenir temporairement des informations (Wikipedia (2021) : https://en.wikipedia.org/wiki/Working_memory). Les études sur la mémoire de travail s'intéressent à la capacité de maintenir des informations et de les manipuler temporairement (voir aussi Baddeley, 1992 ; Cowan, 2001). Selon différentes approches du domaine psychologique, le terme de *mémoire à court terme* est également utilisé. Mais la mémoire à court terme est considérée plutôt comme l'opposé de la mémoire à long terme, et ses études s'intéressent à la capacité que l'on a de maintenir temporairement des informations qui seront oubliées après une certaine durée.

Deuxièmement, un autre phénomène très connu est la limitation du niveau d'auto-enchâssement (angl : *center embedding*) d'une construction syntaxique. Lewis (1996) pense que ce phénomène pourrait être causé par la limitation de la mémoire de travail.

Une construction auto-enchâssée est définie par Miller et Chomsky (1963) comme :

> *"[A] nesting of dependencies, which occurs when a constituent X is embedded in another constituent Y, with material in Y to both the left and right of X"*

Dans l'exemple (2), on trouve une construction enchâssée *X* qui est *that the girl scared* à l'intérieur d'une autre construction *Y* qui est *The dog ran away*, avec du matériel de *Y* à la fois à gauche et à droite de *X.*

(2) [*The dog* [*that the girl scared*]$_X$ *ran away*]$_Y$
Lewis (1996)

Il suffit d'ajouter une seule construction auto-enchâssée de plus pour que la phrase devienne incompréhensible, comme dans (3) :

(3) [*The cat* [*that the bird* [*that the mouse chased*] *scared*] *ran away*].
Lewis (1996)

0.1.2 Défis actuels

L'objectif de cette thèse est de trouver des bons moyens d'évaluer la complexité syntaxique des phrases et de découvrir les limitations de la complexité syntaxique dans les langues. L'étude de la complexité syntaxique présente de nombreux défis, d'autant que les recherches antérieures s'étendent sur plusieurs domaines qui manquent de connexions fortes entre eux.

La psycholinguistique a élaboré diverses théories du traitement de la phrase. Nous présenterons celles qui impliquent le niveau syntaxique : le modèle de la localité de dépendance

(Gibson,1998 ;2000) (voir la section 1.4.2), le modèle fondé sur l'activation (Lewis & Vasishth, 2005) (voir la section 1.4.3.1) et le modèle fondé sur l'attente (Levy, 2008) (voir la section 1.4.3.2).

Les neurolinguistes s'intéressent également au traitement syntaxique des phrases et ont développé des hypothèses à ce sujet en observant l'activité cérébrale (voir la section 1.5), mais les résultats dans ce domaine ne sont pas étroitement connectés avec d'autres domaines.

Dans le domaine de la linguistique, des métriques ont été proposées (voir section 0.2 pour la définition de la métrique) et utilisées pour évaluer la complexité syntaxique dans un corpus (Blache, 2011). La plupart des travaux actuels portent sur la métrique appelée la longueur de dépendance [3] (voir également les sections 0.2.1 pour la définition et 3.3.3.2 pour une description détaillée des travaux liés à cette métrique). Cette thèse utilise également des métriques pour étudier la complexité syntaxique. De plus, nous aimerions nous référer davantage à des études antérieures sur la complexité syntaxique dans divers domaines pour tenter de formuler des hypothèses plus riches et proposer des métriques différentes pour la complexité syntaxique.

Un autre défi est qu'il existe de nombreuses langues dans le monde, et qu'elles présentent de nombreuses différences, notamment au niveau de la syntaxe. Dans cette thèse, nous voulons savoir s'il existe des métriques qui nous permettent d'évaluer la complexité syntaxique des phrases de manière générale. Alors que les études précédentes ont utilisé un petit nombre de langues, nous allons pouvoir mener nos expériences sur plus de 100 langues grâce à la collection de treebanks *Universal Dependencies*. L'objectif est de calculer et d'évaluer la complexité syntaxique d'autant de langues que possible afin de tester nos hypothèses sur la complexité syntaxique (par exemple : est-ce qu'il existe une limite universelle à la complexité syntaxique dans les langues naturelles ?).

Pour répondre au premier défi, nous introduirons successivement en 0.2 la notion de métrique

[3] Les relations de dépendance proviennent des grammaires de dépendance, que nous expliquerons dans la section 2.3.

ainsi que le flux de dépendance, ce dernier nous aidant à élaborer diverses métriques pour étudier la complexité syntaxique. Ensuite, pour répondre au deuxième défi, nous décrirons brièvement dans la section 0.3 le corpus et dans la section 0.4 la méthodologie de recherche utilisés dans cette thèse. Enfin, nous donnerons le cadre structurel de l'ensemble de la thèse dans la section 0.5.

0.2 Métrique

Nous avons mentionné plus haut que nous utiliserons des métriques pour évaluer la complexité syntaxique. Nous donnons donc d'abord la définition de la métrique ici :

Une métrique sur un espace ***X*** **d'unités linguistiques** est une fonction qui associe à ***X*** des valeurs numériques positives ou nulles qui vont constituer une certaine forme de mesure sur ces unités.

0.2.1 Longueur de dépendance

Parmi les métriques de la complexité syntaxique, la plupart des travaux antérieurs se concentrent sur l'étude de la longueur de dépendance (Hudson, 1995 ; Liu, 2008 ; Ferrer-i-Cancho, 2006 ; Gildea & Temperley, 2010 ; Futrell et al, 2015), qui est le nombre de mots entre le gouverneur et son dépendant dans une relation de dépendance. La figure 1 montre un arbre syntaxique, et sous chaque mot, la longueur de la dépendance qui lie ce mot à son gouverneur.

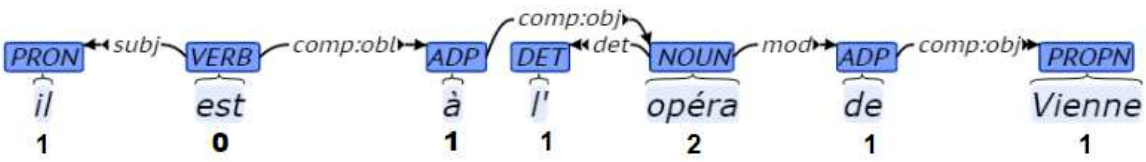

Figure 1 Exemple avec les longueurs de dépendance

Pour une phrase ou un texte, nous pouvons additionner toutes les longueurs de dépendance, ou en faire la moyenne. Cela nous donne une longueur de dépendance totale ou moyenne (voir la section 3.3.3.2 pour plus de détails). Par exemple, pour la figure 1, la longueur de dépendance totale de la phrase est de 7, et la longueur de dépendance moyenne est de 1,17.

Du point de vue cognitif, un mot traité doit être maintenu dans la mémoire de travail, jusqu'à ce qu'il rencontre son dépendant ou son gouverneur. Plus il est maintenu longtemps dans la mémoire de travail, plus il est probable que ce mot soit détérioré ou oublié. La longueur de dépendance est liée à la durée pendant laquelle un mot traité reste dans la mémoire jusqu'à ce qu'il rencontre son dépendant ou son gouverneur. En effet, de nombreux chercheurs ont constaté que les longueurs de dépendance ont tendance à être minimisées dans les langues, ce qui a conduit au développement d'une théorie appelée théorie de minimisation de la longueur de dépendance (Liu, 2008 ; Ferrer-i-Cancho, 2006 ; Gildea & Temperley, 2010 ; Futrell et al, 2015).

0.2.2 Flux d'information

Comme nous l'avons évoqué au tout début, notre hypothèse est que la complexité du traitement des phrases n'est pas seulement liée à la durée pendant laquelle les mots sont maintenus dans la mémoire de travail, mais aussi au nombre d'éléments maintenus simultanément dans la mémoire de travail.

Pour étudier cette hypothèse, nous pouvons d'abord essayer de reconstruire le flux d'informations qui doivent être gardées dans la mémoire de travail lors du traitement de chaque mot. Nous pensons que ces informations sont essentielles lorsque la structure de la phrase est linéarisée en une chaîne, ou que l'auditeur essaye de prédire la structure de la phrase entière.

Le flux d'informations est dynamique, ce qui signifie que les données du flux d'informations changent au fur et à mesure que la phrase est traitée. Nous pouvons évaluer le volume du flux d'informations et des limitations qu'il peut avoir. C'est la motivation pour laquelle le flux de

dépendance est proposé, qui est une modélisation du flux d'information syntaxique. Nous allons le présenter dans la section suivante.

0.2.3 Flux de dépendance

Le flux de dépendance capture l'état du traitement au fur et à mesure que chaque mot de la phrase est traité. Le flux de dépendance dans une position donnée (entre deux mots dans une phrase) est l'ensemble des dépendances qui relient un mot à gauche de cette position à un mot à droite (Kahane, 2001).

Comme on peut le voir sur la figure 2, le flux pour chaque position inter-mot (marqué en lignes verticales) consiste en l'ensemble des dépendances qui passent par sa position. Nous pouvons étudier le flux de dépendance et proposer des calculs à partir du flux de dépendance. Par exemple, le nombre de dépendances dans une position est la taille du flux. Dans la figure 2, la taille du flux est indiquée juste à gauche de chaque ligne verticale qui marque une position inter-mot. Pour la position inter-mot entre *told* et *police*, la taille du flux est de 2, car il y a deux dépendances (*iobj* et *ccomp*) qui passent par cette position.

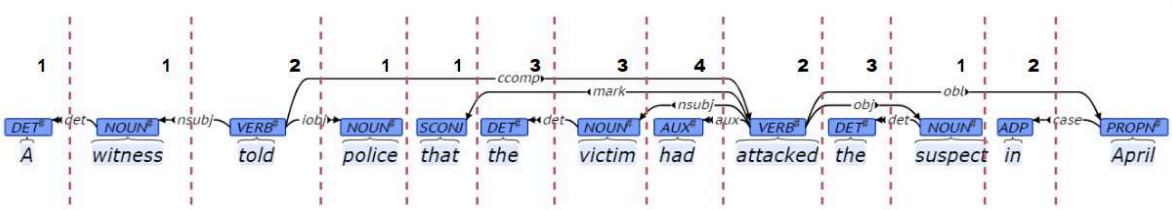

Figure 2 Exemple avec les tailles du flux

Jusqu'à présent, peu de travail sur le flux de dépendance a été effectué. Jardonet (2009) a étudié la taille du flux dans le seul treebank (voir section 2.4 pour la notion et des projets de treebank) du français écrit qui existait à l'époque : *French Treebank* (Abeillé et al., 2003). Il a trouvé qu'il existe des contraintes dans les valeurs de la taille du flux, ce qui est probablement causé par la limitation de la mémoire de travail. Botalla (2014) a étudié le flux de dépendance et les structures non-projectives dans un treebank du français parlé qui s'appelle Rhapsodie (Lacheret

et al., 2014 ; Kahane et al., 2013). L'étude des métriques fondées sur le flux de dépendance est ensuite dévéloppée dans les travaux de Kahane, Yan et Botalla (2017), ainsi que Yan (2017), où nous proposons une métrique que nous appelons le poids de flux : ce dernier permet de capter le niveau de structures auto-enchâssées dans une phrase. Nous avons observé une limitation d'environ 5 pour le poids de flux dans 70 treebanks en 50 langues (voir 4.3.2 et 5.3 pour les détails). Dans cette thèse, nous étendrons les travaux qui ont été faits précédemment sur le flux de dépendance et proposerons de nouvelles métriques du flux fondées sur différentes hypothèses. Enfin, nous effectuerons des analyses quantitatives dans les treebanks.

0.3 Corpus

Cette thèse consiste en des études quantitatives basées sur des corpus. Un corpus est un ensemble de textes qui proviennent de journaux, de magazines, de romans, d'encyclopédies, de documents techniques, de conversations, etc. Sur la base de cette définition, nous pouvons savoir que les données linguistiques contenues dans le corpus sont les ressources attestées. Ce qui diffère du cas de l'utilisation de phrases artificielles dans les études, c'est qu'il existe des phrases artificielles rarement utilisées dans la vie réelle, ce qui entraînerait un biais dans l'expérimentation. Ces dernières années, les données textuelles devenant plus faciles à collecter, nous pensons que les approches de recherche quantitative basées sur des données réelles deviendront de plus en plus fréquentes et fertiles.

Afin de mettre en œuvre nos expériences sur la complexité syntaxique, nous avons besoin de corpus annoté avec des informations syntaxiques. Les treebanks syntaxiques sont composées des phrases annotées en arbre syntaxique (voir aussi section 2.4), ce qui est notre meilleur choix. Étant donné que les treebanks en dépendance sont largement utilisées (par exemple dans les diverses tâches de traitement automatique des langues, telles que l'analyse grammaticale automatique), il existe une importante quantité de projets et ils sont en langues variées. De plus, le flux de dépendance que nous avons introduit plus haut est construit à partir d'annotation en dépendance, il est donc naturel pour nous d'utiliser les treebanks en dépendance.

Actuellement, le projet *Universal Dependencies* (UD) (Nivre et al., 2016 ; 2017) constitue l'ensemble de treebanks en dépendance le plus grand et le plus utilisé (voir aussi section 2.4.1). Ce projet fournit une collection de treebanks dans plus de 100 langues, développées par des équipes du monde entier. De plus, les treebanks UD contiennent des corpus de différents genres, tels que les textes journalistiques, religieux, des conversations téléphoniques et des interviews, etc. Tous ces avantages offrent des conditions très favorables pour nos études.

0.4 Méthode de recherche

Une fois que nous aurons proposé une métrique basée sur le flux, il sera possible d'implémenter des algorithmes, puis d'effectuer des calculs dans des corpus et d'analyser les résultats obtenus. Enfin, à partir de l'analyse des résultats pour la métrique étudiée, nous pourrons vérifier ou contester notre hypothèse de départ.

Dans la section 0.2, nous avons mentionné deux métriques de flux, la taille de flux et le poids de flux, pour lesquelles les chercheurs ont également adopté cette approche expérimentale : calculer la taille de flux ou le poids de flux dans des treebanks et étudier la distribution de leurs valeurs. En examinant la distribution, s'il s'avère que les constructions syntaxiques avec des tailles ou des poids de flux importants sont très rares dans le corpus, alors il est possible que ces constructions soient syntaxiquement plus difficiles à traiter. Si nous constatons que les limites de poids de flux ou de taille de flux de tous les treebanks se situent autour d'une certaine valeur, cela suggère que les locuteurs puissent rencontrer des difficultés à traiter les constructions qui ont des tailles ou poids au-delà de cette valeur. Ils vont favoriser d'autres types de constructions qui ont des valeurs plus petites.

Comme le projet UD comporte un grand nombre de langues, si nous observons que les résultats sont cohérents dans toutes les langues, alors notre conclusion est largement généralisée. Par exemple, il serait très valorisant de découvrir une limite claire de la complexité syntaxique et que cette limite soit universelle pour toutes les langues. Cette universalité pourrait indiquer que

la complexité syntaxique dans les langues naturelles est généralement limitée, et que cette limitation est potentiellement due à la limitation de la mémoire de travail.

0.5 Cadre général

Le premier chapitre présente l'état de l'art de la complexité syntaxique. Il est une introduction panoramique à l'étude de la complexité syntaxique. Ce chapitre fournit les connaissances nécessaires pour lire et comprendre ce qui suit dans cette thèse.

Le deuxième chapitre se focalise sur le corpus annoté en syntaxe. Nous expliquerons tout d'abord pourquoi le corpus est utilisé pour des études de complexité syntaxique. Ensuite, nous introduirons des treebanks en dépendance, car nos expériences dans cette thèse sont réalisées à l'aide des treebanks en dépendance. Enfin, nous présenterons également des classifications des treebanks, car le type de treebank est susceptible d'avoir un impact sur les résultats de l'étude.

Le troisième chapitre présente l'état de l'art sur les métriques de la complexité syntaxique. Dans ce chapitre, nous aborderons des métriques qui ont été proposées pour évaluer d'un point de vue quantitatif la complexité syntaxique et nous exposerons les expériences associées.

Dans le chapitre 4, nous présenterons des métriques fondées sur le flux. Nous donnerons la définition de chaque métrique tout en discutant les hypothèses qui la sous-tendent concernant la complexité syntaxique. Par ailleurs, à la fin de ce chapitre, nous explorerons certaines directions de recherche visant à évaluer si notre métrique reflète la complexité syntaxique dans la réalité.

Enfin, dans le chapitre 5, on présente les résultats de l'étude quantitative de la métrique basée sur les flux. Dans ce chapitre, nous présenterons d'abord des expériences sur la complexité syntaxique des métriques du flux, dans lesquelles nous obtenons des résultats très significatifs. Nous présenterons également dans un deuxième temps d'autres expériences quantitatives, qui

en sont encore au stade exploratoire ainsi que certaines approches dont les conclusions sont similaires aux études précédentes.

1 Etat de l’art sur la complexité syntaxique

1.1. Introduction

Dans ce premier chapitre, nous présentons de façon très générale les principales idées sur la complexité syntaxique, ainsi qu'une vue panoramique des modèles de traitement des phrases. Dans la section 1.2, nous présenterons tout d’abord les études sur les contraintes de la mémoire de travail. Dans la section 1.3, nous déploierons diverses caractéristiques de la complexité syntaxique en nous focalisant sur celles qui serviront à la mise en place des métriques du flux. La section 1.4 présente les modèles de traitement des phrases dans le domaine psycholinguistique. Ces modèles étaieront les hypothèses que nous proposerons pour les métriques fondées sur le flux. Enfin, dans la section 1.5, nous aborderons également des études de la complexité syntaxique dans le domaine neurolinguistique. Cette partie aidera à comprendre en particulier la section 4.4.3 dans le chapitre 4 qui concerne le choix méthodologique d’évaluation des métriques de la complexité syntaxique.

1.2 Contraintes de mémoire de travail

1.2.1 Le nombre magique 7 plus ou moins 2

Le traitement de la structure syntaxique fait appel aux capacités cognitives. Il s’ensuit que les limitations de ces dernières peuvent entraîner des difficultés dans le traitement de structures complexes. Les chercheurs se sont penchés depuis longtemps sur l’étude des limitations de nos capacités cognitives et sur notre potentiel à traiter l’information.

Miller (1956) s’est intéressé aux limitations conséquentes à ce que l’on nomme la mémoire à court terme (angl. *short-term memory*). L'étude sur la mémoire à court terme de Miller (1956) est fondée sur l'observation du nombre d'unités d'information (angl. *information chunk*) et de la quantité d’information (angl. *amount of information/number of bits*) dans chaque unité. Sur

la base d'observations, il a conclu que le nombre d'unités d'information est limité à 7 plus ou moins 2. Il pensait aussi que si nous ne ressentons pas cette limitation au quotidien, c'est parce que nous avons aussi la capacité de recoder l'information en unités plus volumineuses.

1.2.2 La limitation temporelle

Baddeley (1992) a constaté qu'il existe une limitation temporelle dans la mémoire de travail (angl. *working memory*[4]). Selon Baddeley, la mémorisation peut faire recours au processus de répétition (subvocalisée). Sans ce processus, les éléments à mémoriser seront dégradés après 1 ou 2 secondes :

> *"Memory span is a function of both the durability of a trace within the phonological store and the rate at which (subvocal) rehearsal can fresh that trace. If unrehearsed, a phonological trace will fade within 1-2 s. Rehearsal reactivates the trace, and consequently if rehearsal can be repeated every 1-2 s, forgetting will be prevented."*

Cette limite produit un effet dit de longueur de mot (angl. *word length effect*) : les mots plus longs ont un processus de répétition (angl. *rehearsal*) plus long, par conséquent, si la trace n'est pas réactivée, il y aura moins de mots retenus. On observe par exemple ce phénomène lorsque l'on compare la mémorisation des chiffres dans des langues différentes :

> *"The phenomenon accounts for differences in digit span when subjects are tested in different languages; Languages in which digits tend to have long vowel sounds or more than one syllable take longer to rehearse and lead to shorter memory span."*

---

[4] "*Working memory refers to the temporary storage of information in connection with the performance of other cognitive tasks such as reading, problem-solving or learning."* Baddeley (1992). Selon Baddeley, la mémoire de travail contient un centre exécutif, et un composant actif de la boucle phonologique et du carnet à dessin visio-spatial.

1.2.3 La limitation due aux cibles de l'attention

Des études récentes montrent qu'il y a aussi une contrainte de la mémoire de travail due aux cibles de l'attention. Cowan (2001, 2010) se penche sur cette contrainte et pense que :

> *"The fundamental capacity limit appears to coincide with conditions in which the chunks are held in the focus of attention at one time; so it is the focus of attention that appears to be capacity-limited."*

On mesure les contraintes engendrées par les éléments sur lesquels se focalise l'attention grâce au nombre de *chunks* traités simultanément et retenus par l'homme. Cependant, les *chunks* sont difficiles à délimiter. Si à l'intérieur d'un *chunk,* les éléments sont faiblement associés alors la capacité de mémorisation pourrait être sous-estimée. Par ailleurs, les *chunks* qui ont des liens assez forts peuvent se combiner en un plus gros *chunk* et alors on pourrait cette fois surestimer la capacité de mémorisation.

Afin de mieux examiner ces limitations, Cowan (2001) redéfinit la notion de *chunk* comme suit :

> *"A collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use."*

Comme la répétition permet de réactiver l'information aux cibles de l'attention, le processus de répétition doit être désactivé pendant les expériences[5].

> *"The rehearsal could result in a recirculation of information into the focus of attention, reactivating the information each time" (Cowan, 2001).*

[5] Dans les expériences de Chen et Cowan (2009) et de Cowan (2010), pour désactiver le processus de répétition subvocalisée, on demande au sujet de prononcer au cours de la mémorisation d'une liste des *chunks,*en même temps, les mots '*the, the, the ...*'.

Finalement, Cowan (2001) sur ses dernières expériences est arrivé à la conclusion suivante : la capacité de la mémoire de travail, limitée par l'attention, est de 4 *chunks*.

Bien qu'il s'agisse de différents cadres théoriques actuels de la mémoire de travail, les contraintes présentées dans les expériences mettent en évidence un principe fondamental : la capacité de la mémoire de travail est limitée. C'est une condition préalable à l'étude si l'on s'intéresse à la complexité du traitement de phrases ; plusieurs travaux ont déjà montré qu'il existe un lien entre la performance du locuteur (Chomsky, 1965) et la mémoire de travail. Tout d'abord, Baddeley et al. (1985) ont trouvé une corrélation significative entre la mémoire de travail et la compréhension de la lecture chez différents individus ; ensuite, selon Daneman et Carpenter (1983), ce sont les individus qui ont la plus grande mémoire de travail qui ont la compréhension la meilleure des phrases qui contiennent un cas de g*arden-path*[6] (Bever, 1970).

1.3 Phénomènes syntaxiques induisant de la complexité

En général, la difficulté de traitement n'est pas la même pour toutes les phrases. Quand le traitement d'un type de phrase donnée est plus difficile qu'un autre, nous considérerons que ce type de phrase est plus complexe. De même, quand le traitement d'une structure est plus difficile que le traitement d'une autre, nous ferons l'hypothèse que cette structure est plus complexe au niveau syntaxique que l'autre. Dans cette section, nous voulons réfléchir sur les phénomènes de la complexité syntaxique. Nous avons distingué pour cette question de la complexité syntaxique trois points. Dans chaque cas, il s'agit de présenter une hypothèse avec des exemples associés. Le premier cas c'est la complexité de la structure elle-même. Le deuxième c'est la complexité syntaxique en tant que facteur qui influence l'ordre des mots.

[6] Le *garden-path* fait référence aux jardins où un chemin bifurque. Exemple : *The horse raced past the barn fell.* Le destinataire peut faire une analyse erronée avant le mot *fell* : *[The horse raced past the barn]…* où *the horse* est sujet du verbe principal *raced* et *past the barn* est comme un groupe adverbial pour *raced*. L'arrivé de *fell* provoque une réanalyse afin d'établir une bonne structure syntaxique: *The horse [(that was) raced past the barn] fell.*

Enfin, le troisième cas concerne la complexité syntaxique en tant que facteur de désambiguïsation.

1.3.1 Niveau d'auto-enchâssement

Il s'agit dans un premier temps d'une contrainte structurelle pour la phrase. Nous pouvons nous apercevoir de ce qu'il est impossible d'augmenter le niveau d'auto-enchâssement (angl. *center-embedding*) (Miller & Chomsky, 1963 ; Bever, 1970 ; Lewis, 1996 ; Gibson, 2000). La structure d'auto-enchâssement a été définie en 1963 par Miller et Chomsky :

> *"[A] nesting of dependencies, which occurs when a constituent X is embedded in another constituent Y, with material in Y to both the left and right of X" (Example(4)).*

(4) [Y…[X…]…]

(5) Niveau 0. *Le chat miaule.*
Niveau 1. *Le chat* [*que le chien poursuit*] *miaule.*
Niveau 2. *Le chat* [*que le chien* [*que l'homme aime*] *poursuit*] *miaule.*
(Wehrli, 1989)

L'exemple (5) nous montre le niveau d'auto-enchâssement de trois phrases de la langue française. Dans la phrase de niveau 1, *que le chien poursuit* intervient dans la structure *le chat miaule*. En augmentant le niveau d'auto-enchâssement en 2, même si la phrase est grammaticalement correcte, elle est déjà impossible à comprendre. Cette impossibilité serait causée par des contraintes de mémoire de travail (Lewis, 1996).

1.3.2 Rejet des constituants lourds

La complexité de la structure pourrait avoir une influence sur l'ordre des mots. Lorsqu'une structure syntaxique peut avoir plus d'une linéarisation possible, une préférence dans le choix

peut être observée. L'hypothèse est que cette linéarisation préférée serait moins complexe au niveau syntaxique, donc plus facile à traiter.

Par exemple, pour un groupe verbal contenant deux constituants, nous trouvons une préférence à mettre le constituant court avant le constituant long dans les langues à tête initiale. (Hawkins, 2004).

(6) (a). *The gamekeeper looked* [*through his binoculars*] [*into the blue but slightly overcast sky*].

(b). *The gamekeeper looked* [*into the blue but slightly overcast sky*] [*through his binoculars*].

(7) Jeanne a donné [à Pierre] [un livre que j'ai acheté hier].

Dans (6), nous trouvons deux phrases en anglais avec la même structure syntaxique. Il y a une tendance à mettre *through his binoculars* devant *into the blue but slightly overcast sky* comme (a) au lieu de (b), car *into the blue but slightly overcast sky* est un constituant plus long que *through his binoculars.* Cette tendance se retrouve en français dans l'exemple (7), puisqu'*un livre que j'ai acheté hier* est un constituant plus long que *à Pierre*.

Pour les langues à tête finale (voir 2.4.4), comme le japonais, on observe une préférence à placer le constituant long avant le constituant court (Yamashita & Chang, 2001) :

> *"The two experiments confirmed that length directly affected phrasal ordering in production of Japanese: people tended to place long phrases before short ones, consistent with the corpus analyses."*

1.3.3 Ambiguïté de rattachement

Lorsque l'analyse syntaxique d'une phrase n'est pas unique, il y a une préférence pour une des analyses plutôt que pour autre. Dans l'exemple (8), on préfère rattacher *last week* au verbe *went* de la proposition subordonnée plutôt qu'au verbe principal *claimed*.

(8) *John claimed he went to London **last week**.*

John claimed [*he **went** to London **last week***].

Frazier (1979) a proposé la stratégie de *Late-Closure* :

> *"When possible, attach incoming material into the clause currently being parsed."*

En d'autres termes, il est préférable de rattacher l'élément entrant à la proposition la plus proche qui vient d'être traitée si toutefois c'est possible.

Frazier suggère également que l'élément entrant soit de préférence intégré immédiatement dans la structure traitée. Il constate que cette stratégie pourrait avoir un effet d'économie dans l'analyse syntaxique :

> *"The Late-Closure strategy express a preference for incoming material to be associated with material on its left, which has already been received and analysed, rather than with material on its right (...). By allowing incoming material to be structured immediately, Late Closure has the effect of reducing parser's memory load."*

L'exemple (9) nous montre une phrase avec deux analyses possibles en syntaxe de constituance dans la figure 3 : (a) analyse *with a book* comme une modification de *the girl*. En (b), *with a book* est le modificateur du verbe principal *hit*. On observe une préférence pour l'analyse de la phrase selon la structure (a) : Frazier pense que l'attachement entre *with a book* et *hit* est plus économique que *the girl* et *with a book*, ce dernier provoquant plus de nœuds dans le rattachement. Cette remarque concerne l'autre stratégie syntaxique proposé par Frazier (1979), qui est *Minimal Attachment* :

> *"Attach incoming material into the phrase-marker being constructed using the fewest nodes consistent with the well-formedness rules of the language under analysis."*

(9) *Sam hit the girl with a book*. (Exemple de Frazier (1979))

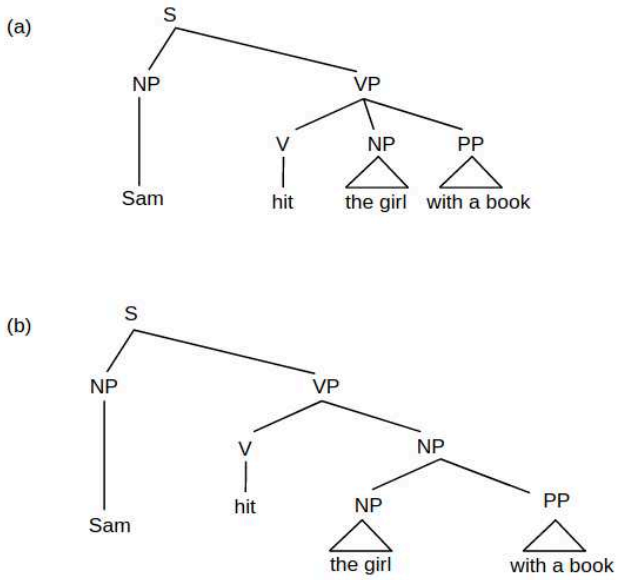

Figure 3 Deux analyses en constituance

Frazier (1979) pense que cette stratégie contribue également à une économie cognitive dans le traitement des phrases :

> *"The Minimal Attachment strategy not only guarantee minimal structure to be held in memory, but also minimizes rule accessing. Hence both Late Closure and Minimal Attachment are 'economical' strategies in the sense that they reduce the computation and memory load of the parser."*

Ces stratégies impliquent qu'il existe des préférences syntaxiques pour la résolution de l'ambiguïté, et que les structures préférentielles seraient moins complexes pour l'analyseur syntaxique.

Aujourd'hui, nous savons que la résolution de l'ambiguïté montrée ci-dessus dépend de plusieurs facteurs et que ce qu'affirme Frazier dépend du modèle syntaxique considéré, comme nous le monterons du point de vue de la compréhension par exemple dans le paragraphe qui suit. Outre la complexité syntaxique, il s'agit aussi des fréquences lexicales, de la plausibilité

et du contexte. En d'autres termes, le choix final de la désambiguïsation ne dépend pas seulement des stratégies de l'analyseur syntaxique que nous venons d'évoquer en haut.

1.3.4 Commentaires

L'explication de ces phénomènes repose sur des arguments liés à la réduction de la complexité. D'un point de vue cognitif, la complexité syntaxique peut jouer différents rôles. Ce qui nous intéresse dans ces phénomènes, c'est le traitement cognitif de la structure, le rôle joué par nos capacités de stockage mémoriel ainsi que nos capacités computationnelles dans la mémoire de travail. Cependant, dans le cas du traitement syntaxique comportant de l'ambiguïté, l'objectif final est de choisir une structure parmi deux structures qui ont des interprétations en sens différentes. C'est la compréhension qui est en jeu. Il est à noter que le traitement d'une phrase en une structure profonde et la compréhension d'une phrase ne relèvent pas du même problème. La compréhension de la phrase est fondée sur les différents niveaux de la cognition. Non seulement les capacités mémorielles, mais aussi les capacités de haut niveau sont nécessaires. En 1.3.3, la difficulté de l'ambiguïté est causée par la présence de plusieurs choix possibles pour prendre la bonne décision considérant le contexte. Afin de parvenir à une compréhension pertinente et adaptée, l'interlocuteur effectue l'analyse sur plusieurs dimensions. En particulier, il faut prendre en compte la dimension pragmatique (King & Just, 1991) et la dimension externe telle que l'analyse de la pensée d'autrui (voir sur la Théorie de la pensée, Duval et al., 2011). La complexité syntaxique fait partie du processus de désambiguïsation, mais d'autres facteurs pourraient renverser la décision dans certains cas. Parfois, le choix final pourrait être une structure plus complexe, mais d'autres éléments peuvent permettre au locuteur de mieux analyser la syntaxe.

1.4 Modèles de traitement des phrases

Actuellement, il existe de nombreuses études sur la complexité syntaxique en

psycholinguistique, et de nombreux développements ont été réalisés. Dans cette section, nous allons d'abord présenter le lien entre la complexité syntaxique et la mémoire de travail. Ensuite, nous nous concentrerons sur les modèles représentatifs du traitement des phrases. Nous commencerons par présenter le modèle de traitement des phrases auquel cette thèse est le plus étroitement liée, à savoir la théorie de la localité de dépendance (angl. *Dependency Locality Theory*) (Gibson 1998 ; 2000). De plus nous évoquerons rapidement deux autres modèles de traitement des phrases importants dans ce domaine, mais accessoires pour notre travail.

1.4.1 Complexité syntaxique et mémoire de travail

Dans cette section, nous allons voir la complexité syntaxique pour deux types de proposition relative : les relatives avec l'extraction du sujet (ES) et les relatives avec l'extraction de l'objet (EO).

(10) (a). **ES :** *The reporter that attacked the senator admitted the error.*
(b). **EO :** *The reporter that the senator attacked admitted the error.*
King et Just (1991)

Dans l'exemple (10), la phrase (a) est une relative avec l'extraction du sujet. Le sujet principal est *reporter,* il représente un rôle de sujet pour le verbe *attacked* dans la proposition relative. Pour la phrase (b), le sujet *reporter* représente un rôle d'objet du verbe *attacked* dans la proposition relative. Selon l'étude de Holmes & O'Regan (1981), les relatives avec l'extraction du sujet sont plus faciles à traiter que celles avec l'extraction de l'objet. Nous nous demandons si la difficulté de traitement de ces deux types de phrases vient de la différence des ressources nécessaires de mémoire de travail. Autrement dit, les phrases relatives avec l'extraction de l'objet sont plus difficiles à traiter, peut-être parce qu'elles nécessitent davantage de ressources en mémoire de travail. Afin de trouver le lien entre le traitement de ces deux types de phrases et la capacité de la mémoire de travail, nous allons présenter une expérience de King et Just (1991).

Expérience 1 de King et Just (1991)

Dans cette expérience, on s'intéresse à trois tests en lecture sur des individus. Tout d'abord, il est nécessaire de tester la capacité de stockage dans la mémoire de travail, c'est-à-dire la possibilité de maintenir les éléments temporairement. Les individus ayant une grande capacité de stockage disposent d'une mémoire de travail plus grande et plus performante. Ensuite, il s'agit de tester la qualité de la compréhension. L'hypothèse est que les individus ayant une faible capacité de mémoire de travail comprennent moins bien les phrases syntaxiquement plus complexes. Enfin, il s'agit de calculer le temps de lecture. En général, une partie d'une phrase lue en un temps plus long est plus difficile à traiter. Par conséquent, le temps de lecture devient un marqueur pour repérer la difficulté de traitement.

Procédure

Les sujets ont fait le test de l'empan de la mémoire de travail et sont répartis en deux groupes : les lecteurs avec empan faible (angl. *low span readers*) (de la mémoire de travail) et les lecteurs avec empan large (angl. *high span readers*). Trois types de séries de phrase(s) sont préparés dans le tableau 1. Chaque série de phrase(s) est présentée devant les sujets (les participants) mot par mot[7]. Après avoir lu une phrase de la série, on demande aux sujets de garder le dernier mot de cette phrase en mémoire. Après la lecture d'une série, il y a deux types de questions auxquelles il faut répondre : premièrement, dire (angl. *recall*) les derniers mots gardés dans la mémoire ; deuxièmement, répondre à des questions pour vérifier si les sujets ont bien compris le sens de la phrase cible.

[7] *"Each set of sentences was presented word-by-word, using a subject-paced moving window paradigm on an IBM-PC/XT"* (Just, Carpenter, & Woolley, 1982).

Type de séries de phrase(s)	Processus
Condition 1 (Une seule phrase)	1. **Phrase cible**[8] : une phrase relative (ES ou EO) ==> Le lecteur garde le dernier mot en mémoire Répondre aux questions
Condition 2 (Deux phrases)	1. Phrase facile à lire[9] ==> Le lecteur garde le dernier mot en mémoire 2. **Phrase cible** : phrase relative (ES ou EO) ==> Le lecteur garde le dernier mot en mémoire Répondre aux questions
Condition 3 (Trois phrases)	1. Phrase facile à lire ==> Le lecteur garde le dernier mot en mémoire 2. Phrase facile à lire ==> Le lecteur garde le dernier mot en mémoire 3. **Phrase cible** : phrase relative (ES ou EO)=> Le lecteur garde le dernier mot en mémoire Répondre aux questions

Tableau 1 Trois types de séries de phrase(s)

[8] Deux exemples de King et Just (1991) :
EO : *The reporter that the senator attacked admitted the error publicly after the hearing.*
ES : *The reporter that attacked the senator admitted the error publicly after the hearing.*
"Each of the sentences ended with an extra prepositional or adverbial phrase that followed the direct object of the main verb. This was done so that the reading time on the direct object was not contaminated by any effect due to encountering the end of the sentence, where subjects sometimes pause briefly (Just & Carpenter, 1980)."

[9] Selon King et Just (1991) : *"The filler sentences that preceded the last sentence in either the experimental or filler sets were sentences that had previously been found to be fairly easy to process, as measured by per-character reading times."*

Différences individuelles de mémoire de travail

Comme dans la procédure d'expérimentation, on demande au sujet de toujours garder en mémoire le dernier mot de la phrase. Cela entraîne une charge mémorielle (angl. *memory load*) supplémentaire. Ainsi, pour les séries à deux phrases, un mot supplémentaire est stocké dans la mémoire de travail pendant le traitement de la phrase cible. Pour les séries à trois phrases, deux mots supplémentaires sont mémorisés pendant le traitement de la phrase cible. Les résultats du rappel des derniers mots ont montré une corrélation avec l'empan de la mémoire chez les différents individus testés. Les sujets ayant une grande mémoire de travail présentent de meilleurs résultats de rappel. En ce qui concerne les résultats des différentes séries, les moins bons résultats sont ceux que l'on obtient sur les séries à trois phrases dont la relative EO est la phrase cible :

> *"The recall rate on three sentence trials was 7% lower if the final sentence was an object rather than a subject relative (...). This suggests that object and subject relative sentences differ in the demands they make on working memory, leaving differential amounts of residual capacity to retain a memory load."*

Compréhension et mémoire de travail

Pour les résultats sur la qualité de la compréhension, dans les séries ayant la phrase cible ES King et Just (1991) n'ont pas trouvé de grande différence par rapport à la mémoire de travail. Par contre, la grande différence se situe au niveau des résultats des séries ayant la phrase cible EO :

> *"**High Span subjects** show a **10% decline** in their comprehension of **object relative** sentences as the load increases from zero to two words, but **no decline** in their comprehension of **subject relative** sentences. **Low span subjects,** who as a group show near **chance-level performance** on object relatives with no memory load, show no decline in their comprehension of object relative sentences as the memory load increases, but do **show a***

> ***decline*** *in their comprehension of* ***the easier subject relative*** *sentences as their processing capacity is increasingly consumed by the demands of storing the memory load."*

Temps de lecture et mémoire de travail

Les résultats sur le temps de lecture sont présentés dans la figure 4 de King et Just (1991). Comme la série à trois phrases montre une mauvaise compréhension et un faible rappel des derniers mots, seules les séries à une et deux phrases sont considérées ici (angl. *set sizes 1 and 2*).

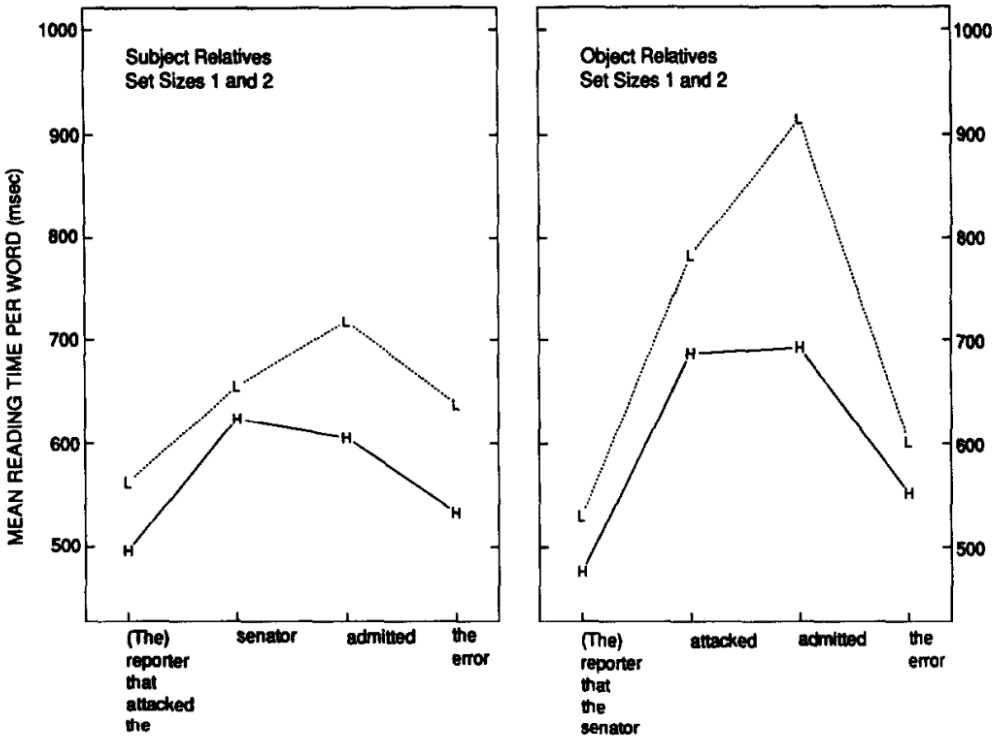

Figure 4. Temps de lecture moyen par mot dans la phrase pour deux groupes d'individus (L: *low span*, H: *high span*), d'après King et Just (1991)

Pour les deux types de phrases relatives (ES et EO), on peut constater que les paritcipants dont la mémoire de travail est supérieure lisent beaucoup plus vite :

> *"There is approximately a 60-ms difference in the reading times of High and Low Span subjects, and both groups of subjects show a similar selective*

increase in reading time at the relative clause-ending sector (senator or attacked) and at the main verb (admitted).”

Par ailleurs, cette différence de temps de lecture entre les deux groupes est plus importante dans la relative EO :

“Although the High Span subjects read more quickly in all four sectors of the object relative sentences, their reading time advantage over the comprehending Low Span subjects is, as predicted, larger in the critical areas and smaller where less processing is required.”

Cette expérience de King et Just (1991) a réussi à montrer qu'il existe un lien entre la complexité syntaxique et la mémoire de travail. Tant le traitement de la structure que la compréhension finale de la phrase nécessitent une mémoire de travail performante. De plus, on pourrait avoir une indication sur la difficulté de traitement par le temps de lecture. Les différents modèles de traitement de phrases sont proposés dans les sections qui suivent. Dans ces modèles, les chercheurs ont essayé de développer des métriques (voir la définition en 0.2). Ils ont aussi conçu des systèmes pour prédire la complexité de la structure syntaxique, nous en parlerons au cours de notre travail.

1.4.2 Théorie de la localité de dépendance

Gibson (2000 ; 1998) a proposé un modèle appelé la Théorie de la localité de dépendance (angl. *Dependency Locality Theory (DLT)*). Cette théorie s’intéresse à l’impact de la mémoire et de la puissance computationnelle sur le traitement de la phrase. Gibson considère que :

“Resources are required for two aspects of language comprehension: 1) the storage of the structure built thus far and 2) integration of the current word into the structure built thus far” (Gibson, 2000).

Les deux aspects mentionnés permettent de développer deux métriques pour prédire la

complexité du traitement des phrases. Les résultats permettent de clarifier les phénomènes de complexité syntaxique, en particulier la différence entre la relative avec l'extraction du sujet (ES) et celle avec l'extraction de l'objet (EO), ainsi que la difficulté à augmenter le niveau d'auto-enchâssement.

Métrique du coût de mémorisation

Selon Gibson (1998), le coût de mémorisation est le nombre de têtes syntaxiques requises pour compléter la chaîne d'entrée en cours d'analyse en tant que phrase grammaticale[10].

Selon la définition ci-dessus, dans l'exemple (11), nous obtenons comme coût de mémorisation de chaque mot les nombres suivants qui sont reportés sous la phrase.

(11) *The reporter who the senator attacked disliked the editor.*

2 1 3 4 3 1 1 1 0

Gibson (2000)

On peut obtenir le coût de mémorisation du premier mot *The* qui est de 2 : il faut au moins un nom pour former le sujet et son verbe pour former une phrase. Comme pour *reporter*, il ne manque qu'un verbe pour former une phrase grammaticale, le coût est de 1. Pour *who*, le coût de mémorisation est de 2, il s'agit de :

> *« **a verb** and **an empty category** position in the relative clause to be associated with relative clause pronoun who » (Gibson, 2000)*

Métrique du coût d'intégration

Dans le contexte de la métrique du coût d'intégration, Gibson introduit deux processus, **le traitement du discours du mot entrant** et **l'intégration structurelle de ce mot**. Il mesure alors le coût en une unité qu'il nomme **Unité d'Energie** du mot (notation **EU** en anglais). Nous

[10] Par exemple, Gibson fait la remarque suivante : "*The minimum number of syntactic head categories in an English sentence is two, a noun for the subject and a verb for the predicate*."

montrons le calcul du coût pour ces deux processus dans le cas présenté dans la figure 5.

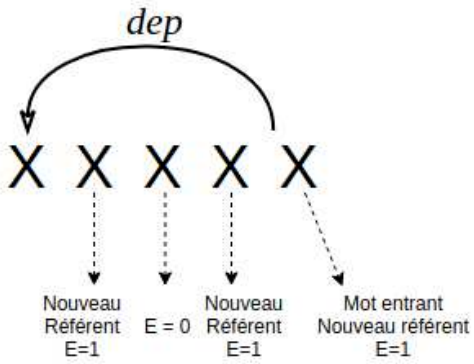

Figure 5. Exemple abstrait pour le calcul du coût d'intégration

Pour le processus du traitement du discours du mot entrant, lorsque le mot entrant introduit un nouveau référent du discours, on trouve en Unité d'Energie un coût de 1 et de 0 dans le cas contraire. Pour le cas de la figure 5, le mot entrant est un nouveau référent du discours, et donc il s'agit d'un coût de 1.

Pour le processus d'intégration structurelle du mot, selon Gibson (2000), nous considérerons que :

> *"The structural integration complexity depends on the distance or locality between the two elements being integrated."*

Pour le cas de la figure 5, le coût de l'intégration structurelle du mot entrant dépend du nombre de nouveaux référents intégrés entre ce mot entrant et le mot déjà traité le plus éloigné auquel il est lié. Si par exemple entre ces deux mots, deux nouveaux référents sont intégrés, on trouve en Unités d'Energie un coût de 2 pour l'intégration structurelle. Ainsi, le coût total d'intégration du mot entrant est de 1 + 2 = 3.

Coût d'intégration dans les relatives

Le coût d'intégration (CI) permet d'expliquer la différence de difficulté entre les phrases relatives avec extraction du sujet (ES) et celles avec extraction de l'objet (EO).

(12) (a). *The **reporter** who _ **sent** the **photographer** to the **editor hoped** for a good **story**.*

0 1 0 1 0 1 0 0 1 4 0 0 0 1

(b). *The **reporter** who the **photographer sent** _ to the **editor hoped** for a good **story**.*

0 1 0 0 1 3 0 0 1 4 0 0 0 1

Dans l'exemple (12), les nouveaux référents sont indiqués en gras. Parmi les 6 nouveaux référents de la phrase (a), il y a 4 nouveaux référents nominaux : *reporter, photographer, editor* et *story*, et 2 nouveaux référents événementiels : *sent* et *hoped*.

On voit que le coût d'intégration du mot *sent* est très différent dans la phrase (a) par rapport à la phrase (b). Pour la phrase (a) qui a une extraction du sujet, quand on traite le mot *sent*, une Unité d'Energie est requise pour la construction d'un nouveau référent événementiel pour *sent,* et il n'y a pas de nouveau référent qui intervienne dans l'intégration structurelle pour le mot *sent*.

Pour la phrase (b), il s'agit de deux intégrations structurelles pour le mot *sent*[11] :

- L'intégration structurelle pour la relation entre *sent* et son sujet *the photographer* ;
- L'intégration structurelle pour la relation entre l'objet vide de *sent* marqué par l'espace vide souligné _ et *who*.

En comparant avec les temps de lecture des participants, on remarque globalement dans les résultats une similarité. En particulier, dans les expériences mesurant le temps de lecture des deux phrases, un temps de lecture plus long est observé dans la phrase (b) par rapport à la phrase (a) pour le mot *sent*. De plus pour les deux phrases, c'est au moment du mot *hoped* que les sujets ralentissent le plus. En conclusion, le coût d''intégration permettrait de prévoir la

[11] *"The NP 'the photographer' is interpreted as the subject and agent of sent (a local integration in which the elements being integrated are linearly adjacent), and the wh-filler who is interpreted as the object and patient of sent (possibly mediated through an empty object). Furthermore, the second of these integrations is a nonlocal integration: the empty object of sent must be integrated back to the wh-filler who across the NP 'the photographer' and the verb 'sent'"* (Gibson & Warren, 2004).

différence de traitement entre ces deux types de relatives.

Coût d'intégration et construction d'auto-enchâssement

Le coût d''intégration a également été utilisé pour prédire la difficulté du traitement dans le cas d'une construction d'auto-enchâssement.

(13) **ES :** *The reporter who the senator* ***attacked*** *disliked the editor.*

CI : **3**

EO : *The reporter who the senator who John met* ***attacked*** *disliked the editor.*

CI : **7**

L'exemple (13) nous montre la phrase (a) avec une seule construction d'auto-enchâssement et la phrase (b) avec elle une construction à double auto-enchâssement. La prévision du coût d'intégration diffère notamment pour le mot *attacked*, 3 Unités d'Energie dans la phrase (a) contre 7 dans la phrase (b). Selon Gibson (2000) :

> *"The cause of processing difficulty associated with nesting complexity is simply that too many long-distance structural integration steps take place at the same point in processing a nested structure."*

Commentaires

En conclusion plus il y a de nouveaux référents entre deux mots qui sont en relation de dépendance, plus ces deux mots seront difficiles à intégrer dans la structure. Ceci permet de développer une métrique formelle pour la complexité syntaxique, appelée la longueur de dépendance. La longueur de dépendance (aussi appelée distance de dépendance) est le nombre de mots entre le dépendant et son gouverneur dans un arbre de dépendance. Elle a été largement reprise et développée dans des études quantitatives ultérieures à propos de la complexité syntaxique, en particulier dans des travaux concernant la tendance dans les langues naturelles à minimiser la longueur de dépendance (Ferrer-i-Cancho, 2006 ; Temperley, 2008 ; Liu, 2008 ; Futrell et al., 2015). Nous reviendrons sur cette métrique au chapitre 3.

1.4.3 Autres modèles

Dans cette section, nous présentons deux autres modèles importants de traitement des phrases. Ils sont proposés un peu plus tard que le modèle de la localité de dépendance de Gibson (2000). Nous présenterons brièvement chaque modèle et discuterons de leurs différences et de leurs liens avec le modèle de Gibson à l'aide d'exemples.

1.4.3.1 Modèle fondé sur l'activation

L'activation est un principe du traitement cognitif dans la théorie *Adaptive Control of Thought-Rational* (*ACT-R*) (Anderson et al., 1998). Le traitement fondé sur l'activation est similaire au mécanisme mis en jeu pour relier le système de production à la mémoire déclarative (angl. *declarative memory*[12]*)*

> *"Different traces in declarative memory have different levels of activation which determine their rates and probabilities of being processed by production rules..."*

Lewis et Vasishth (2005) ont appliqué la théorie d'*ACT-R* à l'analyse syntaxique, et ont développé à ce propos le modèle fondé sur l'activation (angl. *activation based model*). Ce modèle se focalise sur l'interaction de différents aspects cognitifs dans le traitement des phrases. Nous allons voir comment cette interaction entraîne un niveau d'activation différent selon les structures.

Le traitement des phrases est considéré comme un processus qui construit la structure de la phrase de façon incrémentale. C'est ainsi que l'intégration d'un mot entrant dans la structure nécessite ce que l'on nomme la capacité de recherche par indices.

---

[12] *"Declarative memory is one of the two main types of long-term human memory. Declarative memory's counterpart is known as implicit memory (a type being procedural memory), which refers to unconscious memories such as skills (e.g. knowing how to get dressed, eat, drive, ride a bicycle without having to re-learn the skill each time)"* Wikipédia.

Plus précisément, la recherche par indices fonctionne comme suit :

> *"Each incoming word triggers retrievals to integrate that word with the preceding structure. Retrieval is accomplished by a simple type of associative access: content-based retrieval, where the retrieval cues are a subset of the features of the item to be retrieved" (Lewis & Vasishth, 2005).*

(14) *Melissa knew that* ***the toy*** *from her uncle in Bogota* ***arrived*** *today.*
Lewis (2006)

L'exemple (14) nous permet d'expliquer le processus du traitement entre *arrived* et *toy*. Lorsque le traitement se trouve à *toy,* deux informations sont encodées :

- *toy* fait partie d'un groupe nominal, *the toy* ;
- *the toy* en tant que sujet attend son prédicat.

Les processus d'encodage permettent d'activer l'élément entrant et de le stocker temporairement en mémoire. Ainsi, (*the) toy* est activé et sera désactivé s'il n'est pas récupéré plus tard. Quand le traitement se trouve à *arrived*, afin d'intégrer ce verbe dans la structure, un processus de recherche est déclenché. Les indices sont utilisés pour récupérer le mot cible pour mettre à jour la structure. Pour *arrived*, nous pourrions avoir des indices comme :

- *arrived* est un prédicat qui peut former une phrase avec un sujet ;
- ce sujet à récupérer attend toujours son prédicat.

Nous savons déjà que *toy* est un sujet et qu'il est toujours en attente de son prédicat. *Arrived* récupère *toy* pour former une proposition, de sorte que la structure est alors mise à jour.

Le temps de recherche dans la mémoire à partir des indices de l'élément courant dépend du niveau d'activation du mot cible, mais aussi du niveau d'activation de ses concurrents ; ces éléments concurrents sont ceux qui ont un codage sémantique ou syntaxique similaire à celui du mot cible.

> *"An effect of similarity-based retrieval interference is observed if increasing the similarity of distractors[13] (either preceding or following the target) to the retrieval cues used to access the target increases difficulty (and thus reading time or errors) at the point of establishing the relationship that requires the retrieval" (Lewis, 2006).*

Lewis soutient qu'il n'y a pas de surcharge de mémoire, mais plutôt des surcharges dues aux interférences et à la dégradation :

- Plus il y a de candidats similaires dans le processus de recherche, plus il y a d'interférences dans ce processus, et le choix final dépend de leur degré d'activation – c'est le candidat le plus actif qui sera récupéré ;
- Certains mots se dégradent lorsqu'ils ne sont plus actifs, ce qui rend leur rappel plus difficile. De plus, une analyse incorrecte peut entraîner l'échec de l'analyse syntaxique, car certains composants qui auraient dû être activés ne le sont plus et se sont donc détériorés.

Complexité du traitement du verbe dans une construction enchâssée

Le modèle de Lewis propose une explication sur la différence concernant la complexité du traitement du verbe principal et celle pour le verbe de la construction enchâssée (Lewis & Vasishth, 2005 ; McElree et al., 2003).

(15) (a). [sans modification du sujet] *The nurse **supervised** the administrator while…*

(b). [modification prépositionnelle du sujet] *The nurse from the clinic **supervised** the administrator…*

(c). [modification relative du sujet] *The nurse who was from the clinic **supervised** the administrator while…*

(Grodner & Gibson, 2005 ; Lewis & Vasishth, 2005)

[13] Les faux candidats activés dans la mémoire de travail et en attente d'être récupérés.

(16) **Enchâssement:**

*The administrator [who the nurse… **surpervised**] scolded the medic while…*

(d). [sans modification du sujet] *The administrator who the nurse **supervised** scolded the medic while…*

(e). [modification prépositionnelle du sujet] *The administrator who the nurse from the clinic **supervised** scolded the medic while…*

(f). [modification relative du sujet] *The administrator who the nurse who was from the clinic **supervised** scolded the medic while…*

(Grodner & Gibson, 2005 ; Lewis & Vasishth, 2005)

Dans l'exemple (15), (a), (b) et (c) contiennent différents niveaux de modification du sujet du verbe principal *supervised*. Dans (a), il n'y a pas de modification du sujet *the nurse*. Pour le cas de b, le sujet *the nurse* est modifié par un groupe prépositionnel, *from the clinic*. Pour le cas de (c), c'est une proposition relative, *who was from the clinic,* qui le modifie.

Dans l'exemple (16), le verbe *supervised* se trouve toujours dans une proposition enchâssée. (d), (e) et (f) contiennent différents niveaux de modification du sujet *the nurse* du verbe *supervised*.

Le modèle de Lewis et Vasishth (2005) prévoit une augmentation du temps de lecture du verbe *supervised* dans les phrases de (16) par rapport à celles de (15). Cela se vérifie d'ailleurs dans les résultats expérimentaux de Grodner et Gibson (2005). Le temps de lecture augmente également avec le niveau de modification du sujet. Ainsi, le cas le plus difficile à traiter est celui de la phrase (f). Dans (f), il s'agit d'une proposition enchâssée, dans laquelle le sujet de *supervised* est modifié par encore une proposition relative.

Pour *supervised* dans les phrases (d), (e) et (f), ce que l'on recherche c'est son sujet. Cependant, il y a deux groupes nominaux qui sont en attente d'un verbe. *The administrator* attend un verbe comme prédicat de la phrase et *the nurse* un verbe dans la proposition enchâssée. Les deux candidats sont en quelque sorte proches, ce qui provoque une interférence lorsque le verbe cherche son vrai sujet. De plus, pour intégrer *supervised* dans la structure, deux recherches sont

effectuées : le sujet *the nurse* et le gap-filler d'objet *who*. Comme ces deux derniers sont plus éloignés de *supervised*, ils ont dû être maintenus en mémoire plus longtemps que dans les autres exemples. Donc il peut y avoir une dégradation plus grande dans la situation (f) et cela ralentira considérablement le temps de traitement pour *supervised*.

Commentaires

Le modèle de Lewis et Vasishth (2005) comporte des différences avec celui de la théorie de la Localité de dépendance (Gibson, 2000). Premièrement, Lewis et Vasishth (2005) mettent en jeu le degré d'activation et l'interférence. Deuxièmement, le modèle considère qu'il n'y a pas d'ordre des mots dans le processus d'analyse syntaxique. Ce qui est visé là consiste à faire la distinction entre le mot en cours de traitement et les mots passés. La recherche n'est pas fondée sur l'ordre séquentiel comme dans le modèle de la localité de dépendance, mais sur le degré d'activation et d'association entre les deux mots, celui en cours de traitement et celui qui est recherché.

Les travaux de Lewis et Vasishth (2005) nous donnent un aperçu important sur l'étude de la complexité syntaxique. Tout d'abord, nous pensons que l'étude de la complexité syntaxique ne devrait pas être limitée à l'étude de la théorie de la localité de dépendance de Gibson (2000). De plus la complexité syntaxique est liée au type de structure, car le traitement de la structure peut être influencée par le degré d'activation et d'interférence. La complexité du traitement des structures auto-enchâssées sera discutée plus en détail dans les études quantitatives ultérieures (voir la section 4.3.2, et la section 5.3), c'est l'un des points principaux de notre travail. Deuxièmement, les structures auto-enchâssées ont des propriétés internes spécifiques, ce qui peut aussi faire une différence en termes de complexité. Selon le modèle fondé sur l'activation, lorsqu'il y a deux éléments syntaxiques similaires dans le processus de recherche, le locuteur rencontre une difficulté due à des interférences. Pour les structures d'auto-enchâssement, la relation entre sa composition interne et la complexité fait également partie de notre recherche (voir la section 4.3.2 et la section 5.5.1).

1.4.3.2 Modèle fondé sur l'attente

Un autre type d'étude met l'accent sur les attentes contextuelles (angl. *expectation*), comme le modèle fondé sur l'attente (*Surprisal model)* (Hale, 2001 ; Levy, 2008, 2013), que nous allons présenter ci-dessous. Le modèle fondé sur l'attente se focalise en particulier sur la compréhension de la phrase. Pour cette raison, l'hypothèse n'est pas la limitation de la mémoire. Selon le modèle fondé sur l'attente, nous n'attendons pas toujours la fin de la phrase pour commencer à la comprendre. Pendant le traitement du chaque mot entrant, nous avons fait une prédiction à partir de son contexte :

> *"Equally fundamental as the intuition that memory limitations affect online sentence comprehension is the intuition that a language user's context-derived expectations regarding how a sentence may continue can dramatically affect how language comprehension unfolds in real time" (Levy, 2013).*

En ce qui concerne l'interprétation de l'auteur pour la complexité syntaxique, les structures qui correspondent le mieux aux attentes sont celles qui sont les plus faciles à traiter et vice versa.

Valeur de surprise

Le modèle fondé sur l'attente (Hale, 2001 ; Levy, 2008, 2013) formalise le degré de l'attente en considérant son contexte.

La définition de la valeur de surprise est proposée par Hale (2003) :

> *"The **surprisal** of the word in the context it appears—which has raised prospects for a unified treatment of structural ambiguity resolution and prediction-derived processing benefits. **Surprisal** is defined simply as the log of the inverse of the probability of an event."*

F1. $$Surprisal(w_i) = log \frac{1}{P(w_i|CONTEXT)}$$

$$\left[\approx log \frac{1}{P(w_i|w_{1\ldots i-1})}\right]$$

Hale (2003)

La formule F1 exprime que la valeur de surprise du mot dépend de sa probabilité dans le contexte (CONTEXT dans la formule), qui théoriquement inclut à la fois son contexte interne et externe (les facteurs du discours et de la pragmatique). De façon simplifiée, CONTEXT se réduit à ne considérer que le contexte interne (le contexte textuel), qui est composé par la séquence complète du premier mot w_i jusqu'au dernier mot qui précède w_i : (w_1..$w_{i\text{-}1}$).

Prédictions du modèle sur la complexité causée par l'ambiguïté

En particulier, le modèle fondé sur l'attente est justifié pour prédire une difficulté de traitement causée par une ambiguïté temporaire. Nous montrons ici le cas de la construction *garden-path* en anglais.

(17) (a). [*When* [*the dog scratched the vet and his new assistant*]] ***removed*** *the muzzle.*
(b). [*When the dog scratched*] [*the vet and his new assistant* ***removed*** *the muzzle*].

Dans l'exemple (17), (a) et (b) sont deux analyses différentes pour le début d'une phrase avec un *garden-path.* Selon Levy (2013), l'analyse (a) est 5 fois plus probable avant l'arrivée du verbe *removed*[14]. L'ambiguïté est résolue avec le traitement du verbe *removed.* Parce que l'analyse (a) n'est plus conforme au contexte, l'analyse (b) doit être choisie. Cela entraîne un temps de lecture plus lent au niveau de *removed* .

Levy (2013) a également utilisé le modèle fondé sur l'attente pour montrer la difficulté de *removed* dans la phrase ambiguë (18) comparé avec la phrase non ambiguë (19). Levy (2013) a calculé la valeur de surprise de *removed* pour ces deux phrases.

[14] Lewis (2013) : *"We find that the tree in which the vet and his new assistant is interpreted as the direct object of scratched has probability 0.826, and the tree in which it is interpreted as the main clause subject has probability 0.174."*

(18) *When the dog scratched the vet and his new assistant…*

(19) *When the dog scratched its owner the vet and his new assistant…*

D'après les résultats de Levy (2013), la valeur de surprise de (18) est supérieure à celle de (19) (voir Levy (2013) en détail). Une valeur de surprise importante signifie que l'occurrence d'un mot est inattendue : comme le cas dans (18), l'arrivée du verbe *removed* est difficile à prédire à partir de son contexte. Une faible valeur de surprise implique que ce mot est bien prévisible ou facile à prédire, c'est le cas de (19), parce que la non-ambiguïté permet de voir apparaître le verbe *removed* plus sûrement que pour (18).

Relation avec la théorie de la localité de dépendance

Parfois, le modèle fondé sur l'attente obtient des résultats opposés à ceux de la théorie de la localité de dépendance (Gibson, 2000) (section 1.4.2).

(20) (a). *The player* [*that the coach met at 8 o'clock*] ***bought*** *the house…*

(b). *The player* [*that the coach met by the river at 8 o'clock*] ***bought*** *the house…*

(c). *The player* [*that the coach met NEAR THE GYM by the river at 8 o'clock*] ***bought*** *the house…*

L'exemple (20) ci-dessus concerne la prédiction de la complexité du verbe *bought* dans les phrases sans ambiguïté. Avant le verbe *bought*, (a), (b) et (c) contiennent des relatives de longueurs différentes. Dans (a), *that the coach met at 8 o'clock* contient un seul groupe prépositionnel (angl. PP), *at 8 o'clock.* Dans (b), il y a deux groupes prépositionnels : *by the river* et *at 8 o'clock.* Dans (c), trois groupes prépositionnels : *near the gym, by the river* et *at 8 o'clock.*

Selon la figure 6 de Levy (2008), les prédictions du modèle fondé sur l'attente (*Surprisal*) correspondent au temps de lecture réel (*Mean reading time*). Cependant, ce n'est pas le cas pour les prédictions obtenues par la théorie de localité de dépendance (*DLT prediction*).

	Number of PPs intervening between embedded and matrix verb		
	1 PP	2 PPs	3 PPs
DLT prediction	Easier	Harder	Hardest
Surprisal	13.87	13.54	13.40
Mean reading time (ms)	510 ± 34	410 ± 21	394 ± 16

Figure 6 Prédictions du modèle fondé sur l'attente (*Surprisal*) et du modèle de la localité de dépendance (*DLT prediction*), et résultats du temps de lecture, d'après Levy (2008)

Pour la théorie de la localité de dépendance, plus la distance entre *bought* et *player* augmente, plus le coût d'ntégration du mot *bought* augmente. Par conséquent, comme le montre la figure 6, le traitement de *bought* est plus difficile dans la phrase (c), puis pour (b), et dans (a) est le plus facile. Tandis que pour le modèle fondé sur l'attente, (c) est la situation le plus facile :

> *"English subject-modifying relative clauses are a constrained syntactic context. The comprehender knows that the relative clause has to end, but does not know when it will end until seeing the next item of the matrix clause (in this case, the matrix verb). The more post-verbal constituents within the RC (relative clause) that have been seen, the fewer possible choices there are for subsequent constituents within the RC. This follows because constituent types tend to be in complementary distribution – for example, in a given clause the knowledge that a temporal phrase has already appeared makes it less likely that a new temporal phrase will be seen. This means that the comprehender's expectation for the end of the RC (and hence seeing the matrix verb next) should generally increase as the number of already-seen post-verbal constituents increases." (Levy, 2008)*

Cependant, revenons sur l'exemple (15) de Grodner & Gibson (2005). Au verbe *supervised,* le modèle fondé sur l'attente obtient des résultats contraires au temps de lecture et à ceux de la théorie de la localité de dépendance. Dans ce cas, les prédictions de la théorie de la localité de dépendance sont plus pertinentes.

Commentaires

Les travaux de Levy montrent différentes perspectives sur le traitement des phrases. Son modèle nous permet de prédire des valeurs de la complexité plus détaillées que les modèles précédemment exposés, tout au long du traitement des phrases. Comme nous nous appuyons sur la probabilité contextuelle d'un mot, le lexique joue un rôle. Il est possible que deux structures identiques avec des mots différents n'aient pas la même complexité. Enfin, le modèle est spécifique au corpus, puisque l'information sur les probabilités provient du corpus, il n'y a pas d'observation générale au niveau structurel.

Comme nous pouvons le constater à partir d'exemples et d'expériences, l'étude de la complexité du traitement des phrases demande encore beaucoup de travail. En fait, s'appuyer sur l'attente a toujours pour objectif la compréhension de la phrase. Comme nous l'avons exposé précédemment dans 1.2, la compréhension est un processus cognitif de haut niveau qui repose sur plusieurs facteurs. Pour cette raison, il est très utile de résoudre les ambiguïtés.

Dans cette thèse, nous nous intéressons à la complexité syntaxique causée par les limites de la capacité de stockage et la capacité de computation. Notre recherche va consister à proposer des métriques sur les structures syntaxiques rendant compte de ces limites, puis à réaliser des études quantitatives à partir des corpus.

1.5 Neurolinguistique et traitement des phrases

Dans cette section, nous présentons principalement les recherches sur le traitement des phrases dans le domaine de la neurolinguistique ; nous ferons appel à ces notions au chapitre 4. Le domaine de la neurolinguistique mentionne rarement des modèles et des hypothèses de traitement de phrases complètes, bien que les neurolinguistes essaient d'étudier comment le cerveau traite les énoncés linguistiques. À cet égard, il y a principalement deux directions. La première (la section 1.5.1) consiste à examiner les preuves de l'activité neuronale dans le cerveau tout en traitant différents types de phrases. La deuxième direction (section 1.5.2) consiste à utiliser l'information sur l'activité neuronale pour vérifier la prédiction des modèles

de traitement de phrases.

1.5.1 Potentiels évoqués (angl. *event related potentials*)

On peut, pour comparer l'activité du cerveau dans le traitement de différents types de phrases, comparer les données des potentiels évoqués[15] (angl. *event related potentials*). Dans les expériences, nous pouvons obtenir des données sur les potentiels électriques par électroencéphalographie (EEG) et magnétoencéphalographie (MEG). Lorsqu'il y a un événement, les potentiels électriques du cerveau peuvent changer. Les données obtenues par l'EEG ou la MEG sont enregistrées pour chaque participant, la moyenne est calculée pour obtenir un potentiel évoqué. Le rôle du potentiel évoqué dans le domaine du traitement des phrases a été découvert en faisant traiter par les sujets différents types de phrases, telles que des phrases avec des incongruités sémantiques, des phrases grammaticalement incorrectes ou des phrases temporairement ambiguës comme des constructions de *garden-path*.

La N400 et la sémantique

Lorsqu'il s'agit d'une difficulté d'intégration sémantique, une déviation négative peut être observée environ de 400 millisecondes après le stimulus (Kutas & Hillyard, 1984 ; Dambacher et al., 2006). Ce potentiel évoqué est appelé la N400.

(21) (a). *He liked lemon and sugar in his* ***tea****.*
(b). *He liked lemon and sugar in his* ***coffee****.*
(Kutas & Hillyard, 1984)

Dans l'exemple (21), les deux phrases (a) et (b) sont grammaticales. Dans (a), le mot *tea* est très bien congruent avec son contexte. Toutefois, ce n'est pas le cas pour (b). (b) est similaire à (a), sauf que *tea* est remplacé par *coffee*. Même si *coffee* et *tea* sont tous des noms de boissons,

[15] *"A n event-related potential (ERP) is the measured brain response that is the direct result of a specific sensory, cognitive, or motor event."* (Wikipedia (2021) : https://en.wikipedia.org/wiki/Event-related_potential)

coffee est moins probable dans un contexte où il s'agit de *lemon*. Selon les résultats de Kutas et Hillyard (1984), l'amplitude de la N400 est plus marquée après *coffee* dans (b) qu'après *tea* dans (a). De plus, plus le mot remplacé est sémantiquement incongru avec son contexte, plus l'amplitude de la N400 est grande.

L'ELAN, la P600 et la syntaxe

L'ELAN (angl. *early left anterior negativity*) est une déviation négative qui se produit dans le cortex temporal supérieur gauche (angl. *left superior temporal cortex*) entre 100 millisecondes et 200 millisecondes après le stimulus. Le processus lié à ce potentiel évoqué est l'identification de la catégorie syntaxique d'un mot (verbe, nom, préposition, etc.) d'après Friederici (2011 ; 2002). Selon Friederici (2002), l'ELAN est observé lorsqu'il y a une violation de la catégorie syntaxique :

(22) (a). *in a room*
(b). * *in a the*

Dans l'exemple (22) (b), après avoir traité les deux premiers mots, nous constatons que le troisième mot *the* ne correspond pas à l'attente de la catégorie syntaxique. La difficulté se pose lorsque l'on traite *the* pour former un groupe syntaxique en considérant les mots précédents. Ce processus de formation nécessite la connaissance des règles syntaxiques. Comme le mot *the* ne correspond à aucune règle syntaxique pour s'intégrer dans la structure, on détecte une anomalie du traitement syntaxique.

Un autre potentiel évoqué lié au traitement syntaxique est la P600 (la déviation positive d'environ 600 millisecondes après le stimulus). Le potentiel évoqué P600 se retrouve d'abord avec le traitement de la construction *garden-path*, le jugement de grammaticalité (Osterhout & Holcomb, 1992 ; Osterhout et al., 1994). Le processus cognitif correspondant à la P600 n'est pas parfaitement clair, comme le montrent les différents points de vue à ce sujet. Un premier point de vue suggère que la P600 est un potentiel évoqué du processus de réanalyse et de réparation grammaticale (Friederici, 1995 ; Friederici et al., 2002). Un autre point de vue considère, selon Kaan et al. (2000), que la P600 est un potentiel évoqué du processus

d’intégration syntaxique (voir Gibson, 1998, 2000 ; voir aussi 1.4.2). En d'autres termes, lorsqu’il y a une difficulté d'intégration syntaxique du mot entrant, on observe la P600 plus important (Kaan et al., 2000) :

> *“We assume that incoming words and, hence, the syntactic predictions generated by these words are associated with an activation level. This level of activation is a function of distance: when the word cannot be integrated immediately, their activation level decreases as more processing resources have to be devoted to processing new input (e.g., setting up discourse referents for incoming noun phrases, cf., Gibson & Warren, 1997; integration of incoming words, cf., Gibson, in press). When words and their predictions become less activated, integration of current input with these words becomes more difficult: more resources are needed to reactivate these words to allow a successful integration. We claim that the P600 reflects the amount of resources used for these integration processes.”*

L’ELAN et la P600 sont impliqués dans les processus syntaxiques, ceci est différent pour la N400, qui n’est sensible que pour l’intégration sémantique. Selon Hagoort (2003), les potentiels évoqués sont spécifiques à un domaine, mais cela ne signifie pas qu'il n'y a pas d'interaction entre ces domaines. Afin d'étudier les difficultés de compréhension de la phrase, on ne peut pas ignorer la possibilité d’interaction entre les domaines (par exemple, entre la sémantique et la syntaxe). En tout cas, si l’objectif est d’étudier la complexité d'un domaine, les données des potentiels évoqués (comme l'occurrence, l'amplitude, la durée et la latence. (Gouvea et al., 2010)) peuvent fournir des informations très intéressantes.

1.5.2 Localisation fonctionnelle

Un autre type de recherche concerne la localisation des fonctions du langage dans le cerveau. Le travail est effectué en examinant quelles zones du cerveau participent au traitement de différents types de phrases. S'il s'agit de différentes zones de traitement, différents processus

cognitifs peuvent être impliqués. L'expérience la plus courante consiste à examiner l'activité cérébrale à l'aide de l'imagerie par résonance magnétique (IRM).

Aire de Broca

Deux parties du cerveau sont marquées pour le traitement des phrases. L'une est l'aire de Broca, elle est tout d'abord trouvée pour le traitement syntaxique selon les études de Bradley et al. (1980 ; Embick et al., 2000). Selon Caplan et Waters (1999), l'aire de Broca est aussi liée à la fonction de la mémoire de travail. L'autre est l'aire de Wernicke, qui est connue pour le traitement sémantique. Comme l'illustre la figure 7, l'aire de Broca est composée de BA44 et BA 45 et l'aire de Wernicke est composée de BA 42 et BA 22.

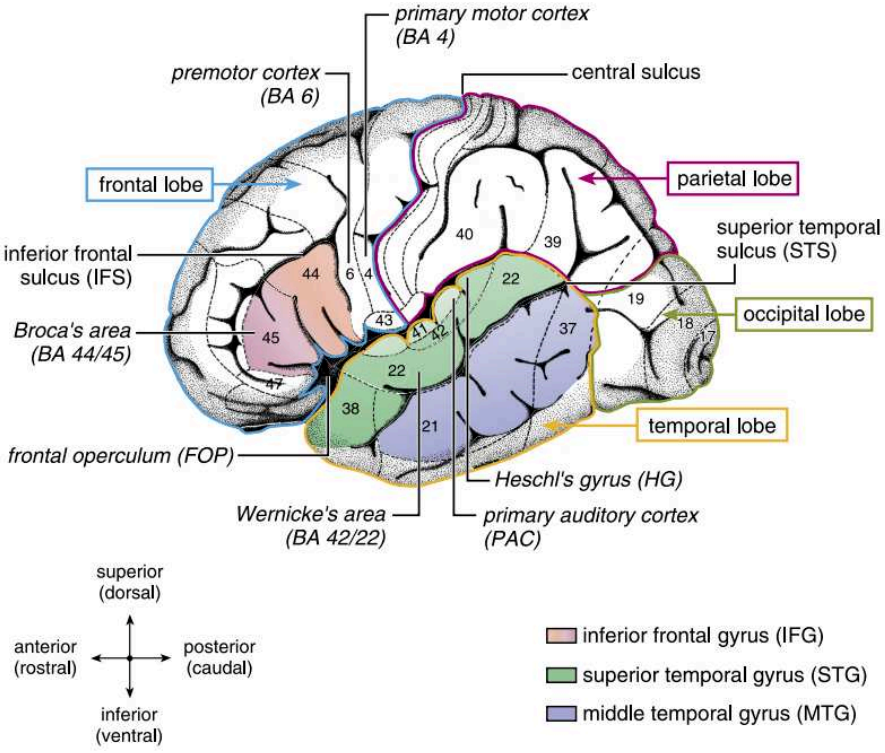

Figure 7 Localisation fonctionnelle d'après Frederici (2011)

Il existe un lien entre le traitement de la syntaxe et l'activation de l'aire de Broca. Comme nous l'avons présenté 1.4.1, le lien entre la mémoire de travail et la complexité syntaxique a été justifié par les expériences de King et Just (1991). Les individus ayant un empan de mémoire de travail plus faible traitent les phrases relatives avec extraction de l'objet plus difficilement.

Rogalsky et al. (2008) ont observé l'activité de l'aire de Broca en testant les relatives avec extraction de l'objet et avec extraction du sujet. Selon Rogalsky et al. (2008), les sous-parties de Broca sont plus activées quand il s'agit d'un traitement de relatives avec extraction de l'objet. Dans l'étude de Friederici (2011), on a trouvé que l'aire de Broca est activée lorsqu'il s'agit de structures auto-enchâssées (angl. *nested structure*) dans la phrase. Cela implique que l'aire de Broca est bien identifiée pour une fonction de traitement syntaxique.

Computation et stockage

L'étude de Makuuchi et al. (2009) s'intéresse à la localisation des zones computationnelles et de la mémoire de stockage dans l'aire de Broca. L'hypothèse est que la longueur d'une relation de dépendance concerne la fonction de stockage, et que le traitement hiérarchique concerne la fonction de computation :

> *"Core syntactic computations are operationalized as processing of the structural hierarchy of a sentence, whereas non-syntactic verbal working memory (VWM) is defined as maintenance cost of the verbal information for a certain period irrespective of the syntactic structure of the sentence."*

L'objectif est de localiser ces deux fonctions dans l'aire de Broca. Deux facteurs sont considérés dans les expériences de Makuuchi et al. (2009) : le niveau d'auto-enchâssement comme facteur de la hiérarchie structurelle et la distance entre le sujet principal et son verbe comme facteur du stockage de la mémoire.

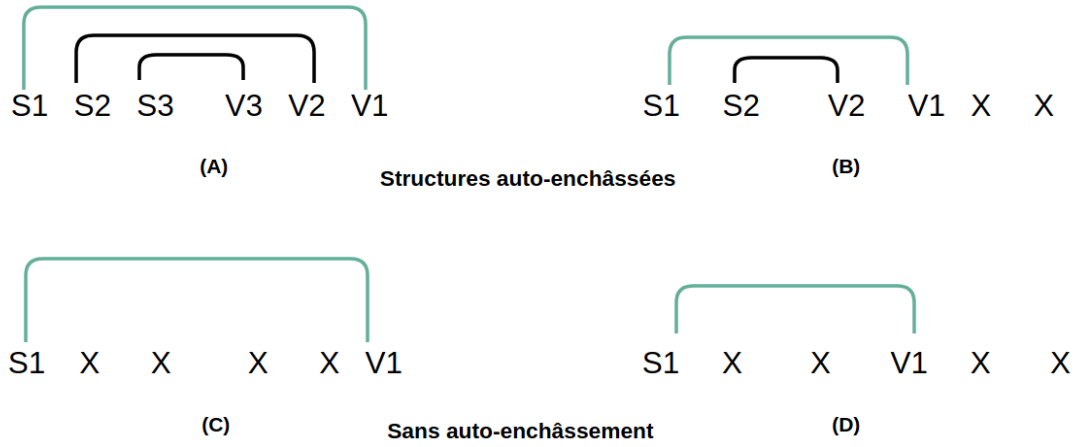

Figure 8 Quatre types de structure : (A) est une construction ayant 3 niveaux d'auto-enchâssement ; (B) est une construction ayant 2 niveaux d'auto-enchâssement ; (C) et (D) sont deux constructions sans auto-enchâssement, ayant une différente longueur de dépendance entre le sujet et le verbe (Makuuchi et al., 2009)

La figure 8 montre des types de structure de phrases en allemand. Les mots sont représentés par S (nom sujet), V (verbe) et X (autre). On s'intéresse à la longueur de dépendance pour chaque configuration entre le sujet principal (S1) et son verbe (V1).

Les configurations (A) et (C) ont la même longueur de dépendance, mais elles ont des hiérarchies différentes, puisque (A) est une structure doublement auto-enchâssée et que (C) est une structure linéaire sans enchâssement. Le même contraste se retrouve entre (B) et (D), où la longueur de dépendance entre S1 et V1 est plus courte qu'en (A) et (C).

En se fondant sur les résultats d'imagerie cérébrale obtenus lors du traitement de différentes structures, Makuuchi et al. (2009) ont constaté que l'activation dans la région BA 44 augmentait avec le niveau d'auto-enchâssement, et que les activités de BA 45 sont affectées par la longueur de dépendance. Les auteurs en déduisent que, bien que ces deux régions soient étroitement liées, elles remplissent des fonctions de traitement syntaxique différentes :

> *"The present data indicate that the cost of keeping track of multiple relations between nouns and verbs in multiply center-embedded structures cannot be accounted for by the distance between dependent items alone; rather it is due to the hierarchical structure, where syntactic computation is indispensable."*

Pour conclure, les données provenant de la localisation fonctionnelle du cerveau peuvent être utilisées pour déterminer les processus cognitifs qui traitent la syntaxe. Makuuchi et al. (2009) pensent que le traitement hiérarchique et le traitement de la longueur de dépendance se fondent sur des capacités cognitives distinctes : la capacité computationnelle et la capacité de stockage.

2 Corpus annoté

2.1 Introduction

Dans ce chapitre, nous nous focalisons sur la présentation du corpus annoté. Tout d'abord, dans la section 2.2, nous expliquerons la raison pour laquelle nous utilisons des corpus annotés pour effectuer nos études. Nous présenterons ensuite dans la section 2.3 le treebank. Nous nous concentrerons sur les treebanks d'*Universal Dependencies* (UD) et de *Surface Universal Dependencies* (SUD), car ces deux projets de treebanks sont les corpus spécifiques utilisés dans nos expériences.

Nous nous intéressons ensuite à la classification des treebanks dans la section 2.4. Les résultats de nos expériences seront influencés par le type de treebanks qui peut être classifié selon diverses dimensions. Dans nos expériences, nous tiendrons compte des différents aspects de la complexité syntaxique concernée pour choisir le type de treebanks le plus adapté. De plus, une partie de notre travail consistera à observer les résultats de nos métriques dans différents types de treebanks.

2.2 Performance linguistique et corpus

Dans son ouvrage publié en 1965, Chomsky fait la distinction entre la compétence linguistique et la performance linguistique. La compétence linguistique est la connaissance de la langue que possède un être humain en tant qu'auditeur ou locuteur. La perception et le traitement du langage par l'homme en tant qu'auditeur ou locuteur sont considérés comme étant la performance linguistique. En d'autres termes, la performance linguistique se concentre sur l'usage de la langue.

Un moyen d'étude de performance est d'analyser l'acceptabilité[16] d'une phrase par l'homme.

L'acceptabilité est différenciée de la grammaticalité, cette dernière étant un sujet d'étude de la compétence linguistique.

Selon Chomsky (1965), le fait qu'une phrase soit grammaticalement correcte ne signifie pas qu'elle est acceptable, cela peut être causé par plusieurs facteurs :

> *"The unacceptable grammatical sentences often cannot be used, for reasons having to do, not with grammar, but rather with memory limitations, intonational and stylistic factors, "iconic" elements of discourse (for example, a tendency to place logical subject and object early rather than late). "*

Dans le cadre du traitement syntaxique, le facteur qui affecte la performance est la mémoire de travail, cette dernière est déjà expliquée dans le chapitre 1. Les structures grammaticalement correctes mais inacceptables sont plus complexes au niveau syntaxique que les acceptables, car elles sont plus contraintes par la limitation de la mémoire de travail. Par exemple, le cas d'auto-enchâssement à niveau 3 décrit à la section 1.3.1 est grammaticalement correct, mais il n'est pas acceptable. Nous pensons que plus les structures sont auto-enchâssées, plus elles sont complexes à traiter, plus elles sont difficilement acceptables.

Le corpus recueille la production réelle écrite ou orale. Toutes les phrases du corpus sont acceptables. Afin d'étudier la performance dans le corpus, on peut formaliser les phénomènes de la complexité syntaxique et faire des études quantitatives. Si un type de structure syntaxique est considéré moins acceptable, elle serait plus rare dans les données réelles, voire impossible à trouver. D'ailleurs, l'avantage d'utiliser le corpus pour l'étude de la performance linguistique est qu'il s'agit d'une méthode indépendante. Dans les expériences psychologiques, les phrases générées artificiellement sont souvent examinées par différents sujets (humains). Parfois,

---

[16] *"The more acceptable sentences are those that are more likely to be produced, more easily understood, less clumsy, and in some sense more natural. The unacceptable sentences one would tend to avoid and replace by more acceptable variants, wherever possible, in actual discourse"* (Chomsky, 1965)

certains sujets jugent une structure acceptable, mais celle-ci n'existe pas en utilisation réelle.

2.3 Corpus annoté

Les études de la complexité syntaxique dans les corpus nécessitent des ressources contenant des analyses syntaxiques. Pour cela, des ressources sous forme de treebanks syntaxiques répondent à nos objectifs. Un treebank est un corpus dont les phrases sont analysées en arbres syntaxiques.

Aujourd'hui, les projets de treebanks sont très développés pour deux raisons. La première est de faire un traitement syntaxique automatique. Avec les méthodes d'apprentissage automatique, les données des treebanks peuvent être utilisées pour entraîner et tester un analyseur automatique (angl. *parser*) (Collin, 2003 ; Nivre, 2008). L'autre raison importante est qu'ils sont des ressources fertiles pour la recherche linguistique. Par exemple, on pourrait étudier une construction syntaxique spécifique en recherchant dans des treebanks ses occurrences ou ses exemples d'exception ; on pourrait également observer les phénomènes grammaticaux d'une langue dans des treebanks ; en ce qui concerne l'étude de la complexité syntaxique, les métriques sont utilisées pour mesurer la complexité syntaxique dans les treebanks.

Il existe actuellement deux approches pour les annotations syntaxiques, l'une basée sur la grammaire de constituants[17] (Bloomfield, 1933 ; Chomsky, 1957) et l'autre sur la grammaire de dépendance[18] (Tesnières, 1959 ; Mel'čuk, 1988 ; Kahane, 2001). Aujourd'hui, les treebanks les plus courants sont annotés en dépendance. Nos études dans cette thèse utiliseront de tels treebanks, c'est-à-dire annotés en dépendance. Nous envisagerons également mais seulement

[17] Wundt (1900) et Bloomfiled (1933) ont été les premiers à proposer l'analyse hiérarchique des phrases en constituants immédiats (angl: *immediate contituent).* L'analyse en constituants immédiats est ensuite reprise par Wells (1947) et Chomsky (1957). Cela a renforcé sa place dans le monde de la recherche linguistique.

[18] La grammaire de dépendance a été proposée dans le livre « Elément de syntaxe structurale » de Lucien Tesnières en 1965. Elle a ensuite été développée et enrichie dans la Théorie de Sens-Texte (voir Mel'čuk (1987) et Kahane (2001)).

brièvement, l'annotation en constituants pour nous focaliser sur l'annotation en dépendance.

Annotation en constituants

L'annotation en constituants se concentre sur les analyses catégorielles et la hiérarchisation des constituants. La figure 9 montre un arbre syntaxique de constituants et la figure 10 son format. Dans la figure 10, la hiérarchie est représentée par l'enchâssement des parenthèses. Nous pouvons voir que la parenthèse la plus extérieure est ROOT. Le deuxième niveau est la racine de l'arbre de constituants S. À partir de cette décomposition par niveau, S est décomposée en NP et VP. Ensuite, NP est décomposé en DT et NN. La parenthèse la plus profonde est le mot avec son étiquette morpho-syntaxique.

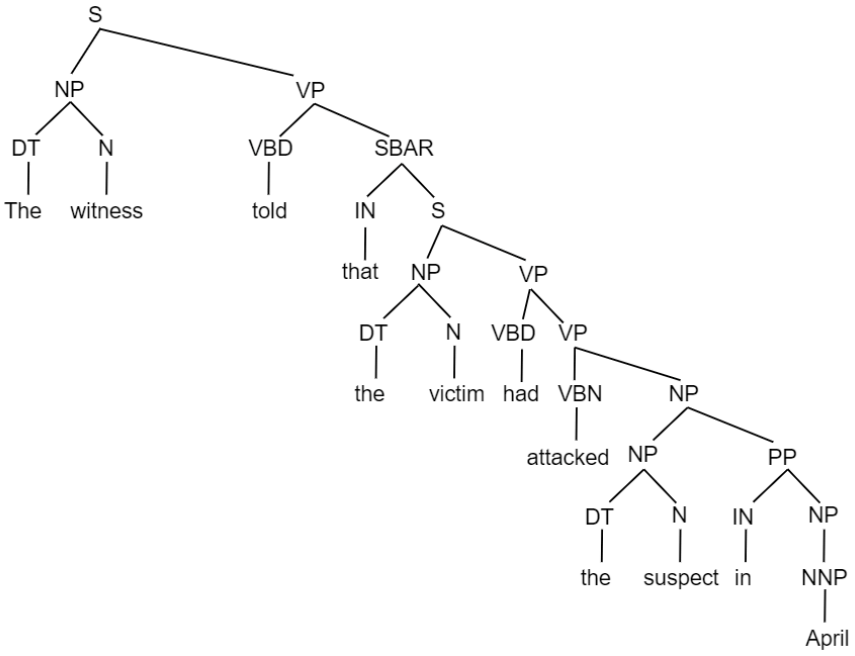

Figure 9 Arbre de constituants

The witness told that the victim had attacked the suspect in April.

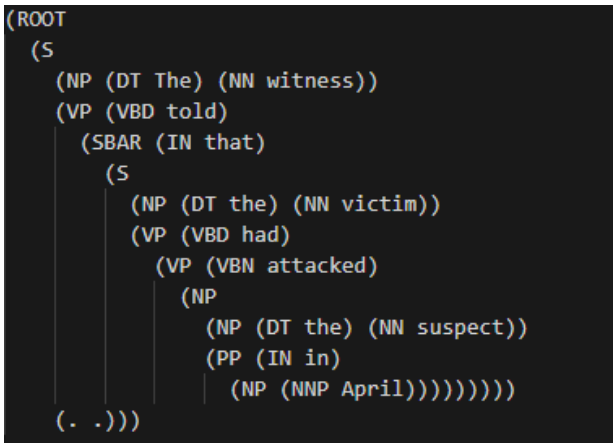

Figure 10 Arbre de constituants encodé dans un fichier texte

Annotation en dépendance

L'annotation en dépendance se concentre sur les analyses des relations, telle que la subordination ou la coordination. Kahane et Gerdes (2021) ont parlé des concepts essentiels concernés :

> *« Il y a **connexion** (syntaxique) dès que deux éléments se combinent pour former un syntagme.*
>
> *…*
>
> *Une connexion hiérarchisée est appelée une dépendance. Si A et B sont connectés et que A est hiérarchiquement supérieur à B, on dit que A **gouverne** B et B **dépend** de A. Ou encore que A est le **gouverneur** de B et que B est une **dépendant** de A »*

Nous pouvons annoter une phrase en arbre de dépendance. Ce qui diffère avec d'autres structures de dépendance c'est que dans l'arbre de dépendance, chaque nœud doit avoir un seul gouverneur.

La figure 11 montre un arbre de dépendance ordonné venant du projet *Universal Dependencies* (voir aussi 2.4.1). C'est-à-dire l'arbre contient non seulement les informations sur des connexions hiérarchisées mais aussi celles sur l'ordre des mots. Toutes les données que nous

utilisons dans les expériences de cette thèse sont de ce type.

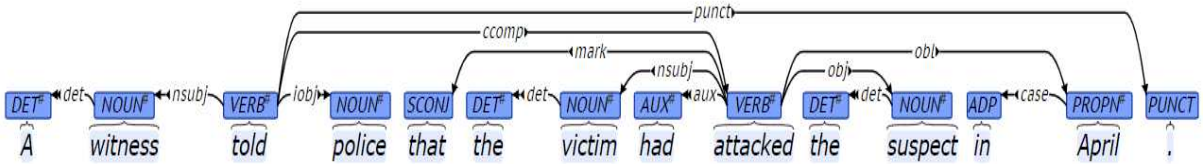

Figure 11 Exemple d'un arbre de dépendance ordonné (UD_English-PUD, n01006011)

Les annotations de l'arbre syntaxique peuvent être encodées dans un format tabulaire dans un fichier texte (la figure 12). Les premières lignes marquées d'un # sont les informations attachées à la phrase elle-même, puis chaque ligne correspond à un mot et aux annotations qui lui sont attachées. Le tableau contient 10 colonnes. Les informations les plus importantes et les plus nécessaires sont la colonne 1 pour l'identifiant du mot, la colonne 2 pour la forme du mot (*token*), la colonne 3 pour son lemme, la colonne 4 pour son étiquette morpho-syntaxique, la colonne 7 pour l'identifiant de son gouverneur (dans le cas de la racine de la phrase, l'identifiant de gouverneur est 0), et la colonne 8 qui donne la relation de dépendance avec le gouverneur (dans le cas de la racine de la phrase, la relation est "*root*"). Les autres colonnes sont remplacées par "_" s'il n'y a pas d'informations à remplir.

```
# newdoc id = n01006
# sent_id = n01006011
# text = A witness told police that the victim had attacked the suspect in April.
1   A          a       DET     DT  Definite=Ind|PronType=Art          2   det     2:det       _
2   witness    witness NOUN    NN  Number=Sing                        3   nsubj   3:nsubj     _
3   told       tell    VERB    VBD Mood=Ind|Tense=Past|VerbForm=Fin   0   root    0:root      _
4   police     police  NOUN    NNS Number=Plur                        3   iobj    3:iobj      _
5   that       that    SCONJ   IN  _                                  9   mark    9:mark      _
6   the        the     DET     DT  Definite=Def|PronType=Art          7   det     7:det       _
7   victim     victim  NOUN    NN  Number=Sing                        9   nsubj   9:nsubj     _
8   had        have    AUX     VBD Mood=Ind|Tense=Past|VerbForm=Fin   9   aux     9:aux       _
9   attacked   attack  VERB    VBN Tense=Past|VerbForm=Part           3   ccomp   3:ccomp     _
10  the        the     DET     DT  Definite=Def|PronType=Art          11  det     11:det      _
11  suspect    suspect NOUN    NN  Number=Sing                        9   obj     9:obj       _
12  in         in      ADP     IN  _                                  13  case    13:case     _
13  April      April   PROPN   NNP Number=Sing                        9   obl     9:obl:in    SpaceAfter=No
14  .          .       PUNCT   .   _                                  3   punct   3:punct     _
```

1	2	3	4	5	6	7	8	9	10
ID	Token	Lemma	POS1		FEATURES	Gov	Relation		

Figure 12 Arbre de dépendance encodé dans un fichier texte

Projets Treebanks

Dans le tableau 2, nous avons énuméré les projets représentatifs de treebanks. Le premier est le projet Penn Treebank (Marcus et al., 1993), qui utilise l'analyse syntaxique en constituants. Le corpus du Penn Treebank de l'anglais contient des articles de presse, des conversations téléphoniques, et des notices techniques, etc.

L'un des premiers treebanks en dépendance est le *Prague Dependency Treebank* (Böhmová et al., 2003) en langue tchèque. Par la suite, de plus en plus de projets de treebanks en dépendance sont apparus. L'un des plus grands projets aujourd'hui est *Universal Dependencies Treebanks* (Nivre et al., 2016). De la version 1.0 à la version 2.7 le projet a inclus 183 treebanks en dépendance en 104 langues provenant d'équipes de recherche du monde entier. Cela constitue une très bonne condition pour calculer la complexité syntaxique dans les arbres des différentes langues. De plus, il y a de nombreux types de textes dans les corpus. Ceci est également important pour l'étude de l'usage des langues selon le genre. Enfin, il y a le projet *Surface Universal Dependencies Treebanks* (appelé SUD), qui consiste à convertir les treebanks d'UD dans une structure de dépendance de surface (basée sur des critères distributionnels), proposé par Gerdes et al (2018). Il s'agit donc de la même source de textes, mais annoté par un schéma d'annotations différentes (voir 2.4.1 sur le schéma d'annotations).

Projet	Langue(s)	Type de texte	Analyse Syntaxique	Référence
Penn Treebanks	Anglais	Journal; Conversations téléphoniques ; Fiches techniques etc.	Constituants	Marcus et al. (1993) Taylor et al. (2003)
Prague Dependency Treebank	Tchèque	Journal;	Dépendance	Hajic et al. (2001) Böhmová et al. 2003
Universal Dependencies Treebanks *(version 1.0 – 2.7)*	104 Langues (Pour la version 2.7)	Journal ; Littérature ; Religieux ; Encyclopédie ; Blog ; Conversations etc.	Dépendance	Nivre et al. (2016) Nivre et al. (2017)
Surface Universal Dependencies Treebanks (version 2.7)	104 Langues (Pour la version 2.7)	Journal ; Littérature ; Religieux ; Encyclopédie ; Blog ; Conversations etc.	Dépendance	Gerdes et al. (2018)

Tableau 2 Projets Treebanks

Nous avons choisi les treebanks en dépendance pour étudier la complexité syntaxique pour deux raisons.

Premièrement, le projet de treebank en dépendance notamment le projet d'UD contient des ressources plus riches pour les différentes langues, ainsi que pour différents types concernant le texte par rapport à ceux en constituants.

Deuxièmement, pour le traitement des phrases, nous avons besoin non seulement d'informations sur la structure syntaxique, mais aussi d'informations sur l'ordre linéaire. L'arbre en dépendance ordonné (voir 3.3.3) contient les deux types d'informations, contrairement à l'arbre en constituants qui contient d'informations linéaires de façon moins explicitée. L'arbre de dépendance ordonné permet d'analyser le traitement intermédiaire. Quant aux arbres de constituants, le principe du traitement consiste plutôt à avoir l'arbre entier et à le linéariser ensuite.

2.4 Treebanks

Cette section se concentre sur certaines classifications des treebanks en dépendance. Tout d'abord, il y a une forme de classification qui est basée sur leurs schémas d'annotation (section 2.4.1 et 2.4.2). D'une part, les schémas d'annotation ont des bases théoriques différentes, ce qui explique pourquoi le choix des tête syntaxiques et l'analyse des structures peuvent varier d'un schéma à l'autre. D'autre part, ils diffèrent également en fonction des objectifs d'utilisation, tels que l'efficacité de l'entraînement de l'analyseur syntaxique automatique ou la recherche de questions typiquement linguistiques. Par conséquent, ces différences affecteront le calcul de la complexité syntaxique. Une autre dimension est le type de texte dans le corpus, aussi appelé le genre du corpus (section 2.4.3). Dans le même schéma d'annotation, les résultats vont varier pour des corpus de genres différents. Finalement, les corpus peuvent également être classés selon les différents types de langues (section 2.4.4), ce qui permet d'étudier la complexité syntaxique dans les langues.

2.4.1 Schéma d'annotation

Dans cette section, nous présentons les deux schémas d'annotations que nous avons déjà évoqués. L'un est le schéma d'*Universal Dependancies (UD)*, et l'autre est celui de *Surface Universal Dependencies (SUD)*.

Les treebanks qu'on utilise dans les expériences de notre thèse sont élaborés à partir de ces deux schémas.

Universal Dependencies

Le schéma d'*Universal Dependencies* (UD) a été développé principalement à partir du schéma d'*Universal Stanford Dependencies* (de Marneffe et al., 2014). Ses étiquettes morpho-syntaxiques se sont basées sur celles de *Google universal part-of-speech tags* (Petrov et al., 2012).

Le schéma UD est un schéma agnostique, selon Nivre (2008) :

> *« Whereas theory-neutral annotation caters for a larger group of users, it runs the risk of not being informative enough or containing too many compromises to be useful for special applications. On the other hand, theory-specific treebanks are clearly more useful for people working within the selected theoretical framework but naturally have a more restricted user group. Recently, there have been attempts at combining the best of both worlds and maximize overall utility in the research community through the use of rich annotation schemes with well-defined conversions to more specific schemes (Nivre 2003; Sasaki et al. 2003). In addition to minimizing the effort required to produce a set of theory-specific treebanks based on the same language data, such a scheme has the advantage of allowing systematic comparisons between different frameworks. »*

Nous trouvons les informations des étiquettes morpho-syntaxiques ainsi que celles des relations syntaxiques sur le site d'UD :

- Les étiquettes morpho-syntaxiques : https://universaldependencies.org/u/pos/index.html
- Les relations syntaxiques : https://universaldependencies.org/u/dep/index.html

Évidemment, étant donné les différences entre les langues, ou les genres, ces informations ne peuvent pas satisfaire tous les phénomènes pour chaque corpus. Ainsi, pour les traits morphosyntaxiques, elles peuvent être représentées dans *Features* (voir colonne 6 de la figure 12, et pour plus d'informations : https://universaldependencies.org/u/feat/index.html) ; pour une relation spécifique tel que *obl:mod* , il s'agit d'une extension *mod* (modifieur) de la dépendance universelle d'origine *obl* (oblique) (https://universaldependencies.org/ext-dep-index.html).

Nous pouvons distinguer deux types de dépendance dans le schéma UD. Le premier est la dépendance entre les mots pleins (angl : *content words*). L'autre est la relation de dépendance vers des mots fonctionnels.

Prenons l'exemple de la figure 13. Les dépendances entre les mots pleins sont : *nsubj*, *iobj*, *ccomp*, *obj*, et *obl ; case*, et *det* impliquent les mots fonctionnels.

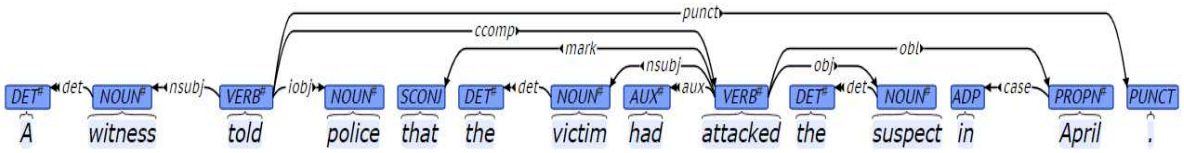

Figure 13 Exemple en anglais (UD_English-PUD, n01006011)

Selon le schéma UD, le mot plein est considéré comme gouverneur du mot fonctionnel dans une relation de dépendance. Cela vise à faciliter la comparaison de corpus en différentes langues, car les relations entre les mots pleins sont plus similaires que celles entre les mots ayant une fonction grammaticale qui varient plus d'une langue à l'autre. Par exemple, la figure 14 montre la même phrase que dans la figure 13 en chinois. Nous pouvons remarquer que dans la phrase en chinois, il n'y a pas de déterminant (relation *det)* ni de marqueur de la proposition subordonnée (relation *mark*), et la relation d'auxiliaire (*aux)* est en direction différente de celle de l'anglais.

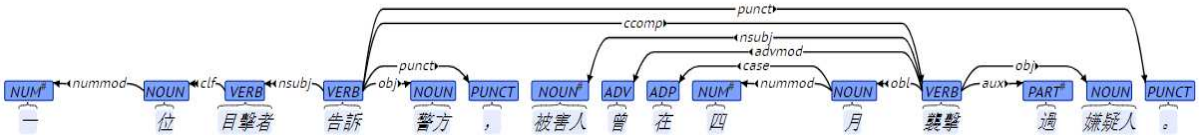

Figure 14 Exemple en chinois[19] (UD_Chinese-PUD, n01006011)

***Surface Syntactic Universal Dependencies* (SUD)**

Gerdes et al. (2018) ont proposé *Surface Syntactic Universal Dependencies* (SUD). Les treebanks SUD sont automatiquement convertis du projet UD à l'aide de règles et peuvent également revenir en UD.

Les étiquettes morpho-syntaxiques du schéma SUD sont identiques à celles du schéma UD. En ce qui concerne les relations syntaxiques, les analyses sont plus étroitement fondées sur les grammaires de dépendance. (voir Tesnière, 1965; Mel'čuk, 1988; Kahane, 2001) :

> *« UD parts with surface syntax criteria and applies the criterion of "content word as head" whereas surface syntax uses distributional criteria of each individual word. The main criterion is that* ***the surface syntactic head determines the distribution of the unit*** *»*

Cela a conduit au fait que l'analyse de la dépendance entre le mot fonctionnel et le mot plein qu'il modifie est très différent de celui d'UD. Étant donné que le mot fonctionnel détermine la distribution du mot plein, il est donc la tête syntaxique et il gouverne le mot plein dans une dépendance. Par ailleurs, il s'agit aussi d'une généralisation de certains types d'étiquettes de relation (voir Gerdes et al. (2018) pour les détails).

---

[19]

一	位		目擊者	告訴	警方，
Yi	wei		mujizhe	gaosu	jingfang
Un	(*classificateur*)		témoin	dire	police

被害人	曾	在	四	月	襲擊	過	嫌疑人。
Beihairen	ceng	zai	si	yue	xiji	guo	xianyiren
victime	autrefois	à	quatre	mois	attaquer	(*PART*)	suspect

Comparaison entre UD et SUD

La figure 15 et la figure 16 montrent la comparaison d'une phrase annotée en UD et en SUD. Nous pouvons constater que les deux schémas donnent lieu à des structures arborescentes différentes.

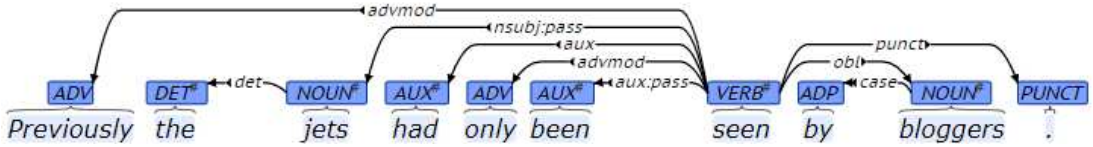

Figure 15 Exemple annoté en UD (UD_English-PUD, n01020004)

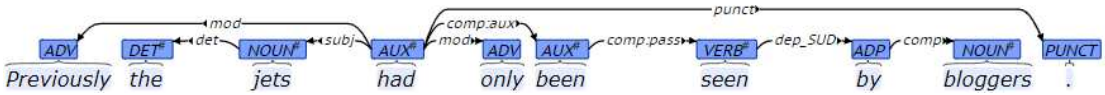

Figure 16 Exemple annoté en SUD (SUD_English-PUD, n01020004)

Comme Gerdes et al. (2018) l'ont souligné, en disposition arborescente, les arbres UD sont plus plats alors que les arbres SUD sont en général plus profonds. Au niveau horizontal, les têtes ont tendance à avoir plus de dépendants dans l'arbre UD : pour l'arbre UD de la figure 15, la tête de la phrase *seen* gouverne 7 dépendants. Pour l'arbre de la figure 16, la tête de la phrase est *had*, et il gouverne 5 dépendants. Au niveau vertical, la différence se manifeste dans le nombre maximal de nœud à partir de la racine jusqu'au nœud terminal. On appelle cela la profondeur (voir 3.3.2.2 et 3.3.2.3, pour la notion de la profondeur). Pour la figure 15, la profondeur est de 2. Par contre, pour l'arbre de la figure 16, la profondeur est de 4.

Ces deux annotations peuvent conduire à des résultats différents, nous allons voir dans nos expériences quelle annotation nous permet de mieux évaluer la complexité syntaxique (voir aussi 3.3.2.3).

2.4.2 Analyses en différentes granularités

En fonction de l'objectif et de l'hypothèse de l'étude, nous pouvons également changer la granularité de l'analyse pour les treebanks. En faisant cela, on génère de nouveaux arbres pour

l'étude quantitative des métriques.

Selon Kahane et Gerdes (2021), la granularité est expliquée comme suit :

> *« La granularité de l'analyse dépend de la taille des unité minimales considérées. Pour une granularité donnée, la structure de connexion contient l'instance minimale de chaque connexion, c'est-à-dire l'instance reliant les plus petites unités considérées pour cette granularité. »*

Ainsi, la première étape pour déterminer la granularité de l'analyse est de spécifier l'unité minimale d'analyse. Par exemple, si nous ne considérons que les relations entre les mots pleins, l'unité minimale d'analyse est un groupe de mots contenant un mot plein. Dans l'étude du flux au chapitre 5 (5.3.3), nous générons un nouveau type d'arbre de dépendance en supprimant les relations fonctionnelles et en agrégeant les mots concernés avec leur gouverneur qui est le mot plein qui les représente.

Dans notre travail, nous utiliserons principalement des treebanks UD, qui ont des structures d'arbre plus faciles à convertir en arbre agrégé, puisque UD traite les mots fonctionnels comme des feuilles de l'arbre. Cela favorise l'étude de la complexité syntaxique dans les arbres où sont agrégés les mots fonctionnels, c'est-à-dire où l'on ne considère que les relations entre les mots pleins.

La figure 17 montre une phrase en schéma UD original et en version agrégée qui considère seulement les relations entre les mots pleins. La version agrégée nous permet de voir comment les groupes des mots s'organisent dans la phrase, et au chapitre 5 nous revenons sur ce point lorsque nous étudierons le flux.

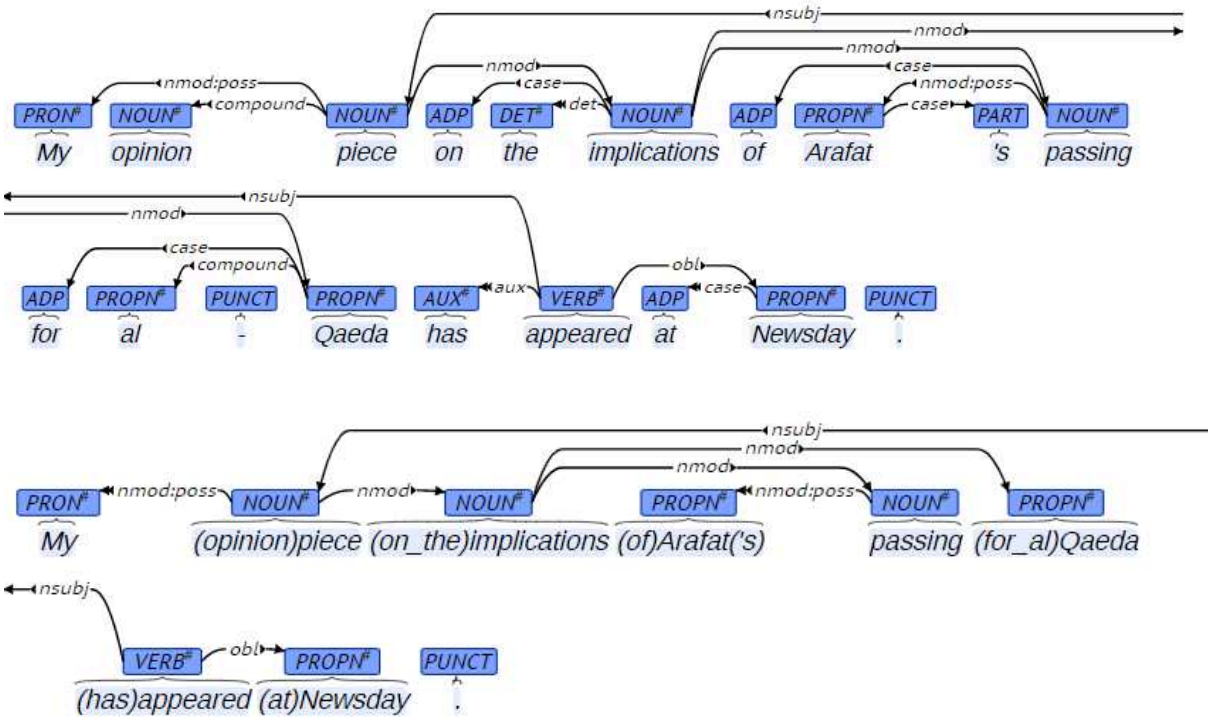

Figure 17 Arbre original et arbre agrégé (UD_English-EWT, weblog-juancole.com_juancole_20041120060600_ENG_20041120_060600-0001)

2.4.3 Langue écrite et langue parlée

Cette section se focalise sur le corpus en langue écrite et en langue parlée. Les travaux sur la complexité syntaxique considèrent principalement deux types qui influencent de fait les résultats des métriques, ce sont celui de la langue écrite et celui de la langue parlée. Dans un premier temps, nous allons expliquer leurs caractéristiques. Par la suite, nous verrons que la langue parlée a des phénomènes syntaxiques différents de ceux de la langue écrite. Nous tenterons de les illustrer brièvement à l'aide d'exemples tirés du corpus français parlé d'UD (*UD_French-Spoken*). Enfin, nous essayerons de présenter les différences de complexité syntaxique entre la langue parlée et la langue écrite.

Caractéristiques

Il est important de noter que le fait que la langue soit parlée ou écrite ne dépend pas du fait qu'elle soit réalisée par le médium phonétique ou graphique. Par exemple, une lettre entre amis

ou SMS de téléphone etc., même s'ils sont réalisés par écrit, ne sont pas des exemples typiques de langage écrit (Koch & Oesterreicher, 2001).

En effet, les concepts de langue écrite et de langue parlée sont plus compliqués à déterminer qu'on ne le pense. Le tableau 3 illustre les propriétés de la langue parlée et de la langue écrite selon différentes approches. La plupart des analyses se concentrent sur le registre de langue. Ainsi, Chafe (1982) et Lakoof (1979) ont considéré que la langue parlée est informelle et que la langue écrite est formelle. L'autre critère concerne la planification ou la préparation (Lakoof, 1979 ; Ochs, 1979). Ce qui distingue dans le discours de la langue écrite de la langue parlée est que le discours dans la langue parlée est souvent spontané et non préparé.

	Langue parlée	**Langue écrite**
Lakoof (1979)	informelle ; spontané	formelle ; préparé
Ochs (1979)	discours non-planifié	discours planifié
Chafe(1982)	Informelle	formelle

Tableau 3 Les propriétés dans la langue parlée et la langue écrite

Koch et Oesterreicher (2001) ont fait la distinction entre la langue parlée et la langue écrite en analysant le comportement communicatif. Le tableau 4 montre les paramètres du comportement communicatif des langues parlée et écrite. Ces caractéristiques sont basées sur la nature de la communication lorsque l'on emploie l'une ou l'autre. La langue parlée possède la nature de l'immédiateté de la communication, tandis que la langue écrite est associée à une forme de la distance dans la communication. Il faut souligner que ces caractéristiques déterminent le genre de texte en continuum :

> *« Le caractère scalaire de l'opposition immédiat/distance est dû premièrement à la gradation internes des paramètres et deuxièmement à la combinatoire des valeurs paramétriques. »*

Langue parlée	Langue écrite
communication privée	communication publique
interlocuteur intime	interlocuteur inconnu
émotionnalité forte	émotionnalité faible
ancrage actionnel et situationnel	détachement actionnel et situationnel
ancrage référentiel dans la situation	détachement référentiel de la situation
coprésence spatio-temporelle	séparation spatio-temporelle
coopération communicative intense	coopération communicative minime
dialogue	monologue
communication spontanée	communication préparée
liberté thématique	fixation thématique

Tableau 4 Langue parlée vs. Langue écrite d'après Koch et Osterreicher (2001)

Ainsi, on ne peut pas simplement supposer que tous les corpus parlés ou tous les corpus écrits ont les mêmes caractéristiques. Cependant, lorsque l'on compare les langues parlées et écrites, on parle aussi de l'universalité de certains phénomènes :

> *« Il est évident que certains phénomènes typiques soit du parlé, soit de l'écrit, se retrouve dans toutes les langues.... Les stratégies communicatives étant déterminée par des facteurs cognitifs fondamentaux, tous ces phénomènes ont un statut universel. » (Koch et Osterreicher, 2001)*

Au niveau syntaxique, certains phénomènes seront illustrés par les exemples de la section qui suit.

L'annotation syntaxique dans les corpus de langue parlée : le cas de *UD_French-Spoken*

Nous allons utiliser le corpus du français parlé d'UD, *UD_French-Spoken*, pour démontrer ses

différences avec la langue française écrite. Ce corpus provient du projet Rhapsodie[20] (Lacheret et al., 2014) et a été automatiquement converti en schéma UD (Gerdes & Kahane, 2017). On présente ici quelques exemples et leur annotation en UD.

La dislocation est très fréquente dans la langue parlée, le constituant disloqué se trouve souvent à gauche, comme dans la figure 18 : *ces rails du tram* est le composant disloqué, comme c'est *rails* qui est la racine de ce composant, il dépend de la racine de la phrase *vais* en relation *dislocated*.

Nous trouvons dans la langue parlée, des marqueurs de discours et les entassements[21] très riches.

La figure 18 contient un type de phénomène d'entassement : la reformulation annotée en relation *conj:dicto*, comme dans le cas de *je vais je vais*. Les marqueurs de discours sont annotés en relation *discourse* et le marqueur de discours dépend du mot à sa gauche (qui lui n'est pas un marqueur de discours).

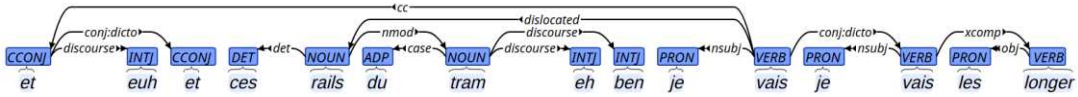

Figure 18 Phénomène d'entassement et marqueur de discours (exemple dans UD_French-Spoken)

et euh et ***ces rails du tram*** *eh ben* ***je vais je vais les longer***

L'exemple de la figure 19 permet de montrer deux autres relations typiques de la langue parlée :

1. *fontaine* - ***parataxis: parenth*** *-> louper*

Dans cet exemple, *vous pouvez pas la louper oui* est une unité illocutoire indépendante[22], qui

[20] Les analyses grammaticales de la langue parlée dans le schéma de Rhapsodie se basent sur les travaux de Blanche-Benveniste (1990 ;2010) et de Deulofeu (2003).

[21] La notion d'entassement couvre la reformulation (annotée en *conj:dicto*), la disflucence (annotée en *conj:dicto*), la coordination (*conj:coord*) et les appositions (annoté en *conj:appos*).

[22] Unité illocutoire (une notion macro-syntaxique) : *"On appelle unité illocutoire une portion de discours comportant un unique acte illocutoire, soit une assertion sur une question"* (Austin, 1975 ; Benzitoun et al., 2010)

est insérée dans une autre unité illocutoire.

2. *là <- **advmod:periph**- allez*

là est un modifieur adverbial non-régi du verbe principal. Il est considéré comme un composant périphérique.

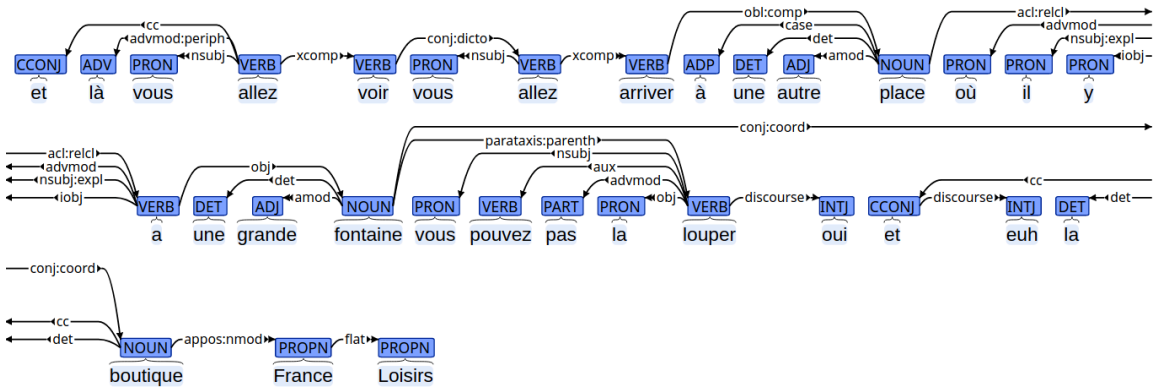

Figure 19 Relation *parataxis:parenth* et relation *advmod:periph* (Exemple dans UD_French-Spoken)

***et** là vous allez voir vous allez arriver à une autre place où il y a une grande fontaine **vous pouvez pas la louper oui** et euh la boutique France Loisirs*

Complexité syntaxique dans langue parlée et écrite

O'donnell (1974) a constaté qu'en anglais parlé, il y avait par rapport à l'anglais écrit moins de constructions de subordination, plus de constructions de coordination, et il pense alors que l'anglais écrit est plus compliqué que l'anglais parlé au niveau syntaxique. Kemper et al. (1989) ont fait des expériences psychologiques, qui les ont conduits à conclure que les structures sont beaucoup moins complexes en anglais parlé tandis qu'à l'écrit on observe une syntaxe moins fragmentée et plus de subordonnées par exemple :

> *« Oral question-answering produced the least complex syntax, whereas written expository statements produced the most complex syntax as indexed*

L'assertion : On peut toujours tester la phrase soit en la niant soit en la confirmant.
Exemple : *Il prend son petit déjeuner*. Test : *Il prend son petit déjeuner ? Oui/Non*.

by an increase in the mean number of words or clauses per utterance, an increase in the use of right- or left-branching embedded or subordinate constructions, and a decrease in the incidence of sentence fragments »

Le comportement communicatif pourrait influencer la performance linguistique. L'immédiateté de la langue parlée conduit au fait que les énoncés ne sont souvent pas planifiés. Sans planification, l'interlocuteur est obligé de traiter une grande quantité d'informations et produire l'énoncé simultanément, ce qui est beaucoup plus difficile. La langue parlée ne favorise donc pas les structures syntaxiquement complexes car la capacité de mémoire est mobilisée principalement ailleurs contrairement à l'écrit qui a la possibilité de présenter dans la phrase de grandes quantités par exemple de situations syntaxiques hiérarchisées dans un même temps.

De plus, on tient à mentionner que le registre de la langue parlée a un caractère informel, car la compréhension peut être facilitée par des facteurs extralinguistiques. Dans le cas de l'écrit, l'expression doit être plus précise, et moins ambiguë. Par conséquent, la langue parlée peut utiliser une syntaxe simple et un vocabulaire moins précis. Par contre, le niveau syntaxique de la langue écrite augmente et le vocabulaire devient plus spécifique.

On peut mesurer la complexité syntaxique des corpus de langue parlée et de langue écrite. Notre hypothèse est que les résultats de la métrique devraient révéler moins de complexité dans la langue parlée et plus de complexité dans la langue écrite (voir le chapitre 5 section 5.3.2).

2.4.4 Corpus en différents types de langues

Le type de langue du corpus est également un aspect à prendre en compte dans les études de la complexité syntaxique. L'un des critères typologiques est la direction de dépendance (Tesnières,1959 ch 14 ; Greenberg 1963). Certaines langues ont tendance à avoir la tête à droite dans une dépendance comme le japonais, l'allemand etc., nous les appelons des *langues à tête finale*. La figure 20 montre un exemple en japonais (*SUD_Japanese-PUD*). On peut constater que toutes les relations (sauf le lien de la ponctuation, non pertinent pour notre discussion) sont

dirigées vers la gauche, c'est-à-dire que le gouverneur se trouve toujours après son dépendant.

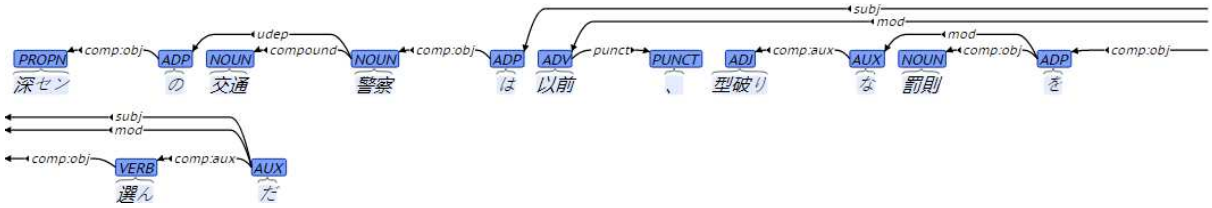

Figure 20 Exemple en japonais (SUD_Japanese-PUD, n01026016)

Shenzhen's traffic police have opted for unconventional penalties before.

Il y a par ailleurs des langues qui ont tendance à avoir des dépendant à droite comme l'irlandais et l'arabe etc. Ce sont des *langues à tête initiale*. La figure 21 montre une phrase en arabe (*SUD_Arabic-PUD*). On peut constater que la majorité des relations sont dirigées vers la droite.

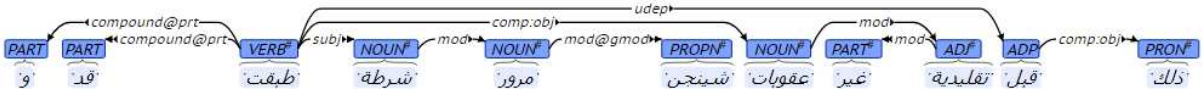

Figure 21 Exemple en arabe (SUD_Arabic-PUD, n01026016)

Shenzhen's traffic police have opted for unconventional penalties before.

Grâce aux travaux sur la typométrie[23] de Gerdes, Kahane et Chen (2021), nous pouvons identifier les langues à tête initiale ainsi que les langues à tête finale. La figure 22 montre les résultats des fréquences de la direction de la dépendance dans les treebanks SUD de Gerdes, Kahane et Chen (2021). Plus une langue tend à mettre le gouverneur devant le dépendant, plus la valeur est grande. Ainsi, plus la valeur pour une langue est proche de 100%, plus la langue a tendance à être de type à tête initiale, alors que plus elle est proche de 0%, plus elle appartient au type à tête finale.

[23] Définition : *"We propose to call **typometrics** the open field of the study of the distribution of languages in a distributional scatter diagram based on empirical measures on corpora. The term **typometrics** is directly inspired by the term **textometrics**, sometimes used for **textual data analysis**. Contrary to textometrics that generally compares texts, authors, or genres from the same language, typometrics attempts to compare languages."* (Gerdes, Kahane et Chen, 2021)

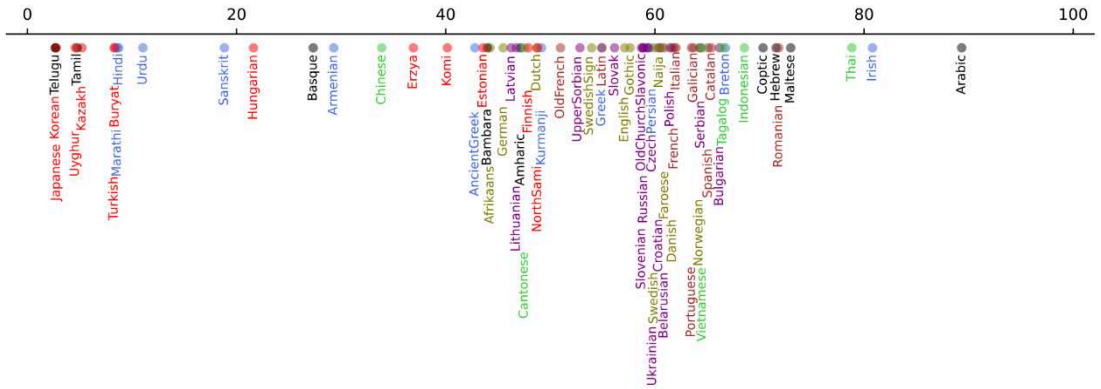

Figure 22 Distribution pour la direction de dépendance dans SUD (Gerdes, Kahane & Chen, 2021)

Nous pensons qu'en même temps que l'on étudie la complexité syntaxique dans des treebanks de différentes langues, il est nécessaire de relier les résultats aux caractéristiques typologiques de la langue. Dans nos expériences sur les métriques du flux, nous pourrons analyser nos résultats en fonction du caractère initial ou final de la position des têtes (voir le chapitre 5).

3 Etat de l'art sur les métriques de complexité syntaxique

3.1 Introduction

Comme nous l'avons évoqué au tout début de cette thèse, une métrique sur un espace ***X*** **d'unités linguistiques** est une fonction qui associe à ***X*** des valeurs numériques positives ou nulles qui vont constituer une certaine forme de mesure sur ces unités. Une métrique est de plus une fonction munie de certains attributs formels que nous allons présenter. Ce chapitre se focalisera sur les métriques de la complexité syntaxique et leurs résultats expérimentaux.

Des premières métriques de la complexité syntaxique ont été proposées dans les travaux d'évaluation de la lisibilité. La section 3.2 donne une brève introduction des métriques utilisées dans ces études, certaines parmi elles ont été utilisés en effet pour la complexité syntaxique. Ensuite, dans la section 3.3, nous explorons les métriques qui permettent de mesurer la complexité syntaxique.

Enfin, dans la section 3.4, nous mettons l'accent sur une contrainte formelle susceptible d'être lié à la complexité syntaxique : la projectivité. Nous présenterons la projectivité, la non-projectivité ainsi que la relation entre la contrainte projective et la complexité syntaxique.

3.2 Mesurer la lisibilité (angl : *readability*)

La lisibilité d'un texte ou d'une phrase fait référence au degré de compréhension d'un groupe de personnes[24]. La lisibilité peut être mesurée au moyen d'expériences psycholinguistiques, dans lesquelles les lecteurs sont invités à effectuer différents types de tests pour déterminer la

[24] Par exemple, les groupes de personnes comme : enfants, adolescents et adultes de différent âge. Ou lecteurs natifs et les apprenants de langues étrangères de différents niveaux.

lisibilité d'un texte. Toutefois, cette méthode est manuelle et requiert beaucoup d'attention pour en contrôler l'objectivité et la rigueur. La deuxième méthode est de proposer des formules d'évaluations : elles consistent en différentes métriques qui s'appliquent notamment au niveau lexical, syntaxique, ou pragmatique. Elles évaluent efficacement la lisibilité de chaque phrase du texte.

Comme le nombre de données linguistiques disponibles est de plus en plus considérable et que les applications informatiques se développent rapidement, on tend à pratiquer cette méthode de manière automatisée et cette approche est maintenant très fréquente. Bien que l'évaluation de la lisibilité n'entre pas dans le cadre de cette thèse, l'idée d'utiliser des métriques pour étudier la complexité linguistique est née de ce domaine. Ainsi, avant de passer aux métriques de la complexité syntaxique proprement dite, nous allons d'abord introduire les métriques de la lisibilité.

Les premiers travaux de l'évaluation de la lisibilité sont ceux de Flesh (1942) et Dale (1948), la longueur des phrases ou des mots a été initialement proposée comme une métrique pour évaluer la lisibilité du texte. Ensuite, il existe également des métriques du niveau lexical. Par exemple, Ure (1971) et Johansson (2008) ont utilisé la densité lexicale (le rapport entre le nombre de mots pleins et la longueur de phrase ou de texte), considérant que les mots grammaticaux ne reflètent pas la complexité de la lecture. Un autre exemple est la variation lexicale (Lieven, 1978 ; Richard & David, 1997), qui est le rapport entre le nombre de mots différents et le nombre total de mots, ce qui suggère que plus la variation lexicale du texte est importante, plus il y a d'information à traiter.

Des études récentes ont utilisé des modèles statistiques pour mesurer la lisibilité du texte. Par exemple, Collins-Thompson et al. (2005) ont utilisé des probabilités d'apparition lexicales pour prédire la difficulté des textes. L'hypothèse est que plus un mot est rare, plus il est difficile à traiter. Ainsi, la difficulté du texte dépend de la moyenne des probabilités d'occurence de tous les uni-grammes (éventuellement aussi bigrammes, ou trigrammes) du texte.

La plupart des métriques utilisées dans l'évaluation de la lisibilité sont au niveau lexical, mais

nous pensons que des informations plus riches sur la complexité syntaxique peuvent également contribuer à l'évaluation de la lisibilité. Comme nous l'avons mentionné au chapitre 1, le traitement des phrases requiert deux types de capacité cognitive : la mémoire joue un rôle clé dans le traitement des structures syntaxiques, et les facteurs lexicaux, sémantiques et environnementaux peuvent permettre la compréhension finale des phrases. Les structures qui dépassent les contraintes mémorielles sont difficiles à comprendre, même avec l'aide des autres facteurs. Nous nous intéressons aux métriques de la structure syntaxique, nous laissons donc de côté les travaux sur les métriques concernant les facteurs lexicaux, sémantiques et contextuels. Dans les sections suivantes, nous allons donc présenter différentes approches pour les métriques de la complexité syntaxique.

3.3 Métriques pour la complexité syntaxique

Il y a dans la cohérence de ce que nous avons exposé au paragraphe précédent trois types de données sur lesquelles nous allons envisager des métriques : les données brutes, les données organisées en arbre non ordonné, et en arbre ordonné. Comme une métrique va dépendre de ces propriétés formelles, on la choisira selon le type de données analysées. De ce fait, la présentation de cette section sera faite selon ces types : section 3.3.1 pour les données brutes, section 3.3.2 pour les arbres non ordonnés et section 3.3.3 pour les arbres ordonnés. Les données les plus fondamentales sont celles en texte brut, celles qui sont les plus accessibles. Les métriques de la complexité syntaxique pour ces données sont la longueur des phrases et le nombre des lettres. Les données en arbre non-ordonné sont constituées de phrases annotées sous forme d'arbre syntaxique, donc des informations structurelles sont fournies. Enfin, les données en arbre ordonné contiennent non seulement des informations structurelles sous forme d'arbre, mais aussi des informations sur l'ordre linéaire. Les treebanks de dépendance d'UD et de SUD sont sous ces formes.

Concernant les commentaires possibles sur chaque métrique présentée, plusieurs critères sont à considérer pour déterminer si elle est appropriée pour mesurer la complexité syntaxique.

Premièrement, une métrique permet d'évaluer la difficulté syntaxique de la phrase, mais aussi celle du traitement de chaque mot de la phrase. Deuxièmement, on verra si et comment elle permet d'interpréter le processus du traitement des phrases. Enfin, la métrique est commode et économique à appliquer. Et dans la pratique l'application d'une métrique dépend aussi de l'efficacité computationnelle et de l'accessibilité des données.

3.3.1 Les données brutes

Longueur de phrase

Pour étudier la complexité syntaxique dans les données brutes, la longueur de la phrase est souvent utilisée. L'hypothèse de départ est qu'une phrase plus longue est plus difficile à traiter qu'une phrase plus courte. La longueur de la phrase pourrait être une approche pour la complexité syntaxique car on constate qu'elle est souvent corrélée avec d'autres métriques de la complexité syntaxique : par exemple, on montre une corrélation entre la longueur de phrase et la longueur de dépendance moyenne (Jiang & Liu, 2015 ; Courtin & Yan, 2019) (voir aussi section 3.3.3.2 pour la longueur de dépendance). En outre, les longues phrases permettent également une plus forte hiérarchisation (Courtin & Yan, 2019), ce qui augmente la complexité syntaxique (voir aussi section 3.3.2.2 et 3.3.2.3 pour les métriques sur la hiérarchisation).

De fait pour diverses raisons on ne retient pas vraiment la longueur de la phrase comme base d'une métrique appropriée de la complexité syntaxique. Premièrement, il nous manque une hypothèse psychologique, la relation entre la longueur de la phrase et la difficulté de la phrase n'est pas causale. Deuxièmement, nous pouvons avoir la même longueur de phrase dans des structures différentes. Il n'est pas possible de comparer la complexité syntaxique de deux phrases de même longueur sans analyse syntaxique. Troisièmement, la longueur de la phrase est une métrique pour la phrase, et ne peut pas fournir une valeur pour le traitement de chaque mot de la phrase. Enfin, il est difficile de définir une phrase : les frontières des phrases dans les données brutes sont souvent déterminées par la ponctuation, mais certains textes bruts comme les conversations ou les entretiens recueillis par des chercheurs de différents domaines, sont

des transcriptions de l'oral qui ne contiennent pas de ponctuation. Par ailleurs, l'utilisation de la ponctuation est aussi influencée par le style et le genre du texte. On pourrait peut-être envisager une approche de la complexité à partir de la présence de signes de ponctuation mais nous n'aborderons pas ce sujet dans notre travail car notre choix a été d'enlever tous les signes de ponctuation dans nos corpus.

3.3.2 Arbre non ordonné

Nous pouvons utiliser des informations sur la structure syntaxique pour analyser la complexité. Les graphes utilisés pour représenter les structures syntaxiques peuvent prendre de nombreuses formes de graphes, l'une d'elles est la forme d'arbre[25]. Actuellement, la plupart des corpus avec analyse syntaxique existent sous la forme de treebanks (voir aussi 2.4). Les arbres syntaxiques par eux-mêmes ne rapportent pas d'information sur l'ordre des mots. Nous définissons donc plus précisément ce type d'arbre comme un arbre syntaxique non ordonné.

La métrique fondée sur la structure arborescente est dépendante de l'approche d'analyse de l'annotation. Ces analyses sont souvent fondées sur divers formalismes, ou sur différents cadres théoriques. D'un point de vue formel, il s'agit soit d'arbres de dépendance soit d'arbres de constituants (voir 2.3). Considérant les cadres théoriques différents, il existe de plus différents types d'arbres de constituants et différents types d'arbre de dépendance : nous constaterons que certains auteurs utilisent les métriques syntaxiques basées sur l'arbre de constituants binaire (voir 3.3.2.1), et que d'autres auteurs utilisent l'arbre de constituants plat (voir 3.3.2.2) ; en ce qui concerne les arbres de dépendance, l'analyse des têtes conduit à choisir des principes différents pour déterminer les dépendances (voir sections 2.4.1 et 3.3.3.2). Toutes ces différences influenceront le choix de la métrique et ses résultats dans les expériences.

[25] Nous rappelons la notion de l'arbre syntaxique : *« Un arbre est un cas particulier de graphe orienté connexe acyclique pour lequel chaque nœud est la cible d'une seule arête à l'exception d'un nœud appelé la racine de l'arbre. »* (Kahane & Gerdes, 2021)

3.3.2.1 Profondeur de Yngve (1960)

Yngve (1960) est un des premiers à avoir proposé des métriques formelles sur les arbres syntaxiques. L'hypothèse de Yngve est qu'il existe une limitation de stockage mémoriel pour générer les règles grammaticales de la phrase.

> *"If the set of sentences that the grammar generates is infinite, there is the possibility that an infinite amount of temporary storage may be required. But a device with an infinite amount of temporary storage is not physically realizable, and we suspect that the human memory, too, does not have an infinite amount of storage capacity"*

Yngve a proposé la profondeur afin de mesurer la capacité de stockage mémoriel. Il pense que cette métrique pourrait avoir une valeur maximale.

Il a analysé la structure syntaxique sous forme d'arbre de constituants binaire. Chaque constituant peut avoir deux enfants non-terminaux[26] : par exemple, le constituant NP (groupe nominal) a les sous-constituants T (déterminant) et N (nom), et le constituant VP (groupe verbal) a sous-constituants V(verbe) et NP (groupe nominal).

La profondeur selon Yngve est calculée comme dans la figure 23 : pour chaque niveau dans l'arbre de constituant, on numérote la branche droite 0 et la branche gauche 1.

Dans le cas de S = NP+VP, la branche de VP est 0, et celle de NP est 1. Ensuite, pour chaque mot, on fait la somme de toutes les valeurs de la branche à partir du mot jusqu'à la racine S. Ainsi, pour le mot *The*, sa valeur de profondeur est $D = 0 + 1 + 1 = 2$. La profondeur maximale D_{max} pour cette phrase est la valeur maximale parmi tous les D, qui est de 2.

[26] L'enfant gauche et l'enfant droit est en fonction de l'ordre linéaire de la phrase. C'est-à-dire que l'enfant gauche devrait précéder de l'enfant droite dans la phrase. Dans ce sens, l'analyse de Yngve n'est pas purement en arbre syntaxique non-ordonnée, car la détermination des enfants non-terminaux a besoin l'information linéaire.

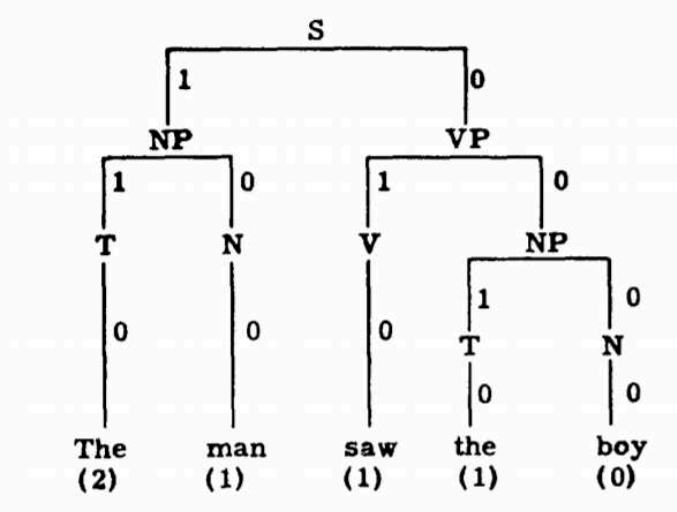

Figure 23 Phrase annotée en arbre de constituants avec les profondeurs d'après Ygnve (1960)

Des études quantitatives qui utilisent la métrique de Yngve (1960) ou une variante de cette métrique (Murata et al., 2001) ont toutes montré une limitation de la profondeur.

Murata et al (2001) ont calculé la profondeur de Yngve (1960) dans le Penn Treebank (comme dans ce dernier corpus, les branches ne sont pas binaires, afin de respecter la numérotation binaire, pour un nœud ayant plus de deux branches, seule la branche la plus à droite est numérotée 0 et les autres sont numérotées 1, voir la figure 24) et ils ont trouvé que la valeur maximale était de 7.

Figure 24 Numérotation des branches pour les arbres dans Penn Treebank d'après Murata et al. (2001)

L'autre variante de profondeur de Yngve (1960) de Murata et al. (2001) consiste à considérer le groupe nominal comme une unité indivisible, il devient donc un nœud terminal. Pour cette nouvelle métrique, seulement 0,01% des mots avaient une profondeur égale ou supérieure à 9 dans *Penn Treebank*. Ils pensent que la profondeur de Yngve soutient l'hypothèse de Miller (1956) sur la limitation de la mémoire à court terme à 7 ± 2 (voir le Chapitre 1, section 1.2).

3.3.2.2 Complexité syntaxique et profondeur de Köhler et Altmann (2000)

Köhler et Altmann (2000) ont proposé également deux métriques syntaxiques pour l'arbre de constituants : **complexité syntaxique** (angl. *syntacitc complexity*) et **profondeur** (niveau d'auto-enchâssement).

Ces deux métriques de Köhler et Altamnn sont proposées à partir de l'arbre de constituants plat. Ce dernier est un type d'arbre de constituants différent de celui utilisé par Yngve (1960).

La figure 25 montre un exemple d'arbre de constituants plat. On peut constater que la racine de l'arbre S est composée de cinq nœuds : NP (groupe nominal), AP (groupe adverbiale), V(verbe), PP (groupe prépositionnel) et SF (phrase finie). Du point de vue d'analyse en dépendance, à part la tête de S qui est le V (*said*), les nœuds NP, AP, PP et SF sont dominés[27] par la tête V (*said*), c'est-à-dire qu'ils sont des groupes de mots dépendant de V.

Si nous regardons tous les autres nœuds, nous pouvons avoir les mêmes observations, par exemple : les constituants P (*in*) et NP (*term-end presentements*) forment un groupe maximal PP, le nœud NP est dominé par la tête P (*in*). Dans ce cas, ce groupe maximal PP est appelé **constituent majeur** de la tête P.

[27] « *Nous dirons qu'un nœud B est **dominé** par le nœud A si B = A ou bien si B dépend de A ou bien si B dépend d'un dépendant de A et ainsi de suite.* » (Kahane & Gerdes, 2021)

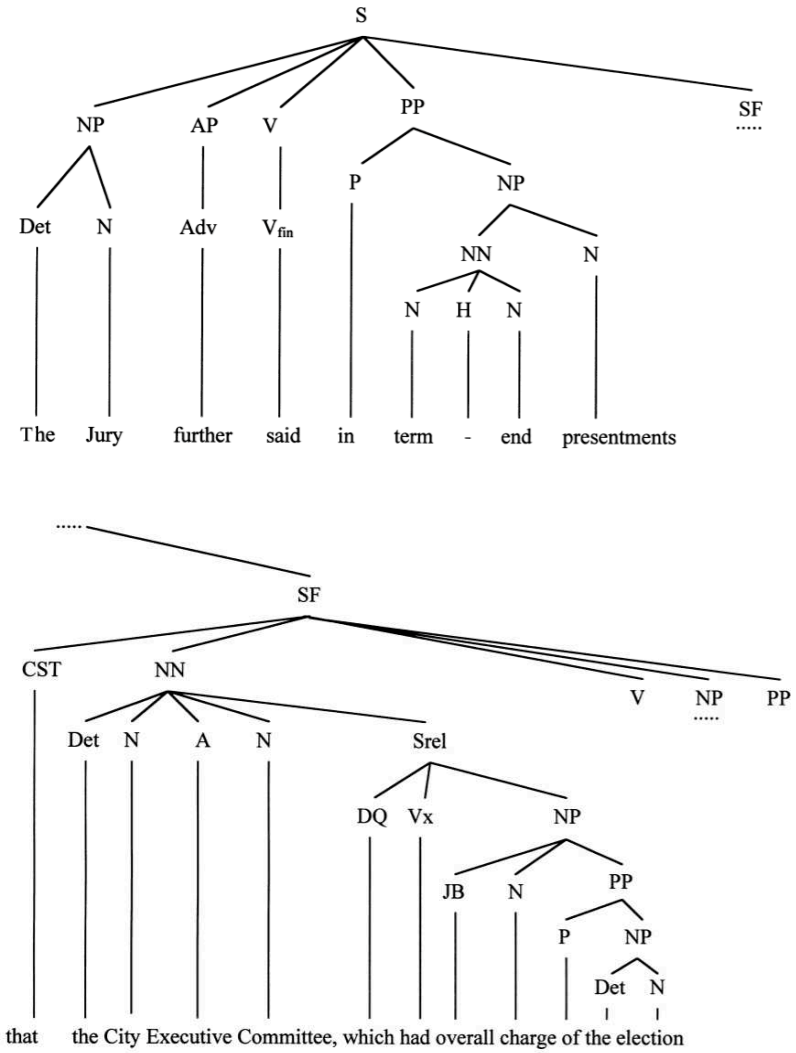

Figure 25 Un arbre de constituants plat utilisé par Köhler et Altmann (2000)

Ainsi, nous pouvons trouver que l'arbre de constituants plat (voir aussi Kahane & Gerdes, 2021) représente un emboîtement de constituants majeurs. Le constituant majeur est le plus grand constituant tel que tous ses nœuds soient dominés par la tête.

Nous savons que la connexion (voir section 2.3) est entre les mots dans l'arbre de dépendance, et par ailleurs l'arbre de constituant plat est une représentation hiérarchique qui montre la connexion entre les groupes des mots. Si nous faisons l'analyse en dépendance, l'analyse en constituants majeurs de S se transforme en Figure 26 : V est gouverneur de N (nom), Adv (adverbe), P (préposition) et V' (verbe de la proposition subordonnée).

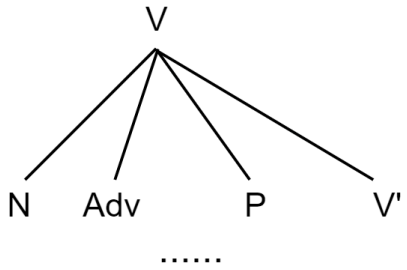

Figure 26 Analyse en dépendance

La complexité syntaxique (Köhler & Altmann, 2000)

La complexité syntaxique pour un nœud est le nombre de sous-constituants qui le compose :

> *« The complexity of a syntactic construction is defined here in terms of the number of its immediate constituents » (Köhler & Altmann, 2000)*

Pour l'arbre de constituants de la figure 25, la complexité syntaxique de S est de 5, car il est composée de 5 sous-constituants :

> NP (The jury) + AP (further) + V(said) + PP (in term-end presentments) +SF (that the City Executive Committee, which had overall charge of the election)

Cette métrique est équivalente à la métrique de l'arité pour l'arbre de dépendance. Cette dernière est définie comme le nombre de dépendants d'un nœud dans l'arbre de dépendance (Courtin & Yan, 2019). Dans l'analyse en dépendance, la tête V, nœud supérieur à ses dépendants (voir la figure 26), doit être remonté. Ainsi, l'arité pour le nœud V est le nombre de ses dépendants qui est de 4. Conséquemment, on constate que la valeur de la complexité syntaxique de Kohler et Altmann est l'arité dans son arbre de dépendance augmentée de 1.

La profondeur (Köhler & Altmann, 2000)

La profondeur proposée par Köhler et Altmann est différente de celle de Yngve (1960). Il s'agit pour chaque nœud de son niveau dans la hiérarchie :

> *« The term of depth of embedding (or briefly: depth) is operationalized as the number of steps from S to the given constituent, by applying production*

rules. »

Dans la figure 25, pour la racine S, la profondeur est 0, dans le cas du nœud NP (*the jury*), la profondeur est de 1, et 2 pour le NP (*term-end presentments*).

La profondeur de Köhler et Altmann est équivalente à la distance hiérarchique de l'arbre de dépendance de Jing et Liu (2015). Cette dernière sera définie comme la distance verticale entre un nœud et la racine. Nous allons en reparler dans la section 3.3.2.3.

Distribution

Köhler et Altmann (2000) ont analysé la distribution des valeurs de la complexité syntaxique et de la profondeur dans les corpus (corpus Susanne[28] et corpus Negra[29]). Ils ont trouvé que les distributions de leurs valeurs suivent la loi hypergéométrique[30]. La figure 27 (Köhler & Altmann, 2000) montre la distribution de la complexité syntaxique dans le corpus Susanne et la figure 28 (Köhler &Altmann, 2000) celle de la profondeur. F[x] est en colonne verte représente l'occurrence observée de complexité syntaxique et NP[x] est l'occurrence estimée par la fonction hypergéométrique (Köhler et Altmann ont utilisé le terme : *hyper-pascal*). On peut constater que tous les deux ont tendance à monter au début, et se réduisent progressivement lorsque leur valeur est plus grande. Ainsi, les valeurs des deux métriques ont des contraintes, ce qui serait possiblement causé par l'effort de la mémoire.

[28] Corpus SUSANNE : corpus en anglais, voir Sampson (2002)

[29] Corpus NEGRA : corpus en allemand, voir Skut et al. (1997), Brants et al. (1999)

[30] *Fitting* une loi : dans la tradition de la linguistique quantitative, à partir des observations sur les données, on fait une hypothèse sur la distribution des valeurs d'une métrique. Et le *fitting* est la recherche du meilleur modèle probabiliste qui corresponde aux données observées.

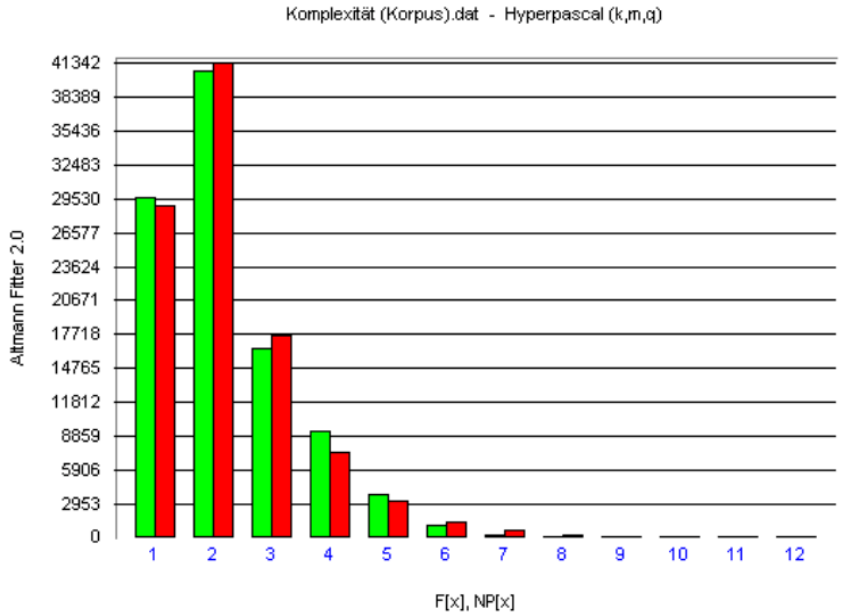

Figure 27 Distribution de la complexité syntaxique (corpus Susanne) d'après Köhler et Altmann (2000)

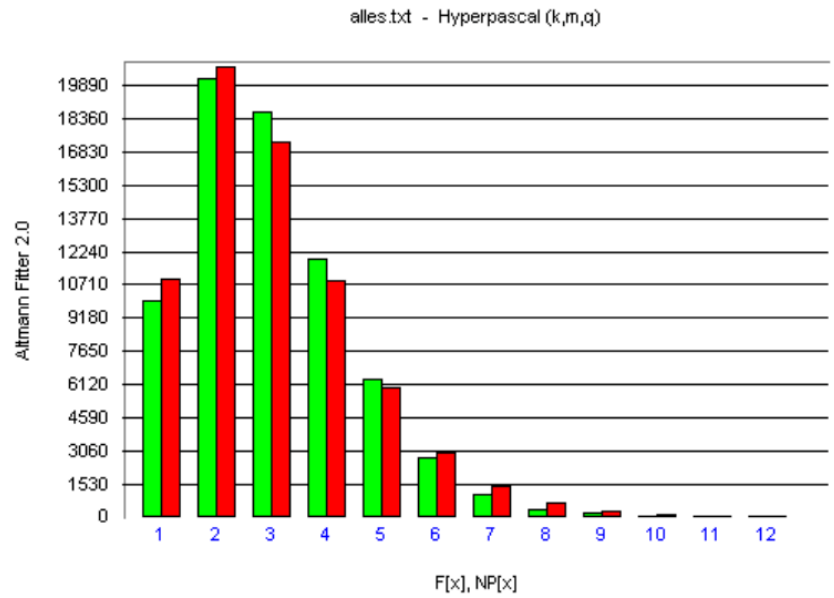

Figure 28 Distribution de la profondeur (corpus Susanne) d'après Köhler et Altmann (2000)

Les études ultérieures qui sont faites par Yang (2019) ont analysé la profondeur de Köhler & Altmann (2000) dans cinq langues (allemand, anglais, chinois, portugais, et japonais), Yang a également trouvé une limitation pour sa valeur et les distributions se modélisent bien toutes avec la loi hypergéométrique pour les cinq langues. Cette distribution en hypergéométrique pourrait être universelle dans les langues (Yang, 2019).

3.3.2.3 Distance hiérarchique

On peut également mesurer la hiérarchie dans l'arbre de dépendance. Jing et Liu (2015) ont proposé la métrique qui s'appelle **distance hiérarchique** (angl. *hierarchical distance*) :

> *"The vertical distance between a node and the root, or the path length traveling from the root to a certain node along the dependency edges, is defined as hierarchical distance"*

Selon cette définition, pour le cas de la figure 29, la longueur hiérarchique moyenne (angl : *Mean Hierarhical Distance*, MHD) est de (2+1+1+2+ 3+3) /6 =1.17. La distance hiérarchique maximale d'une phrase est la hauteur de l'arbre. La hauteur de l'arbre de la figure 29 est de 3.

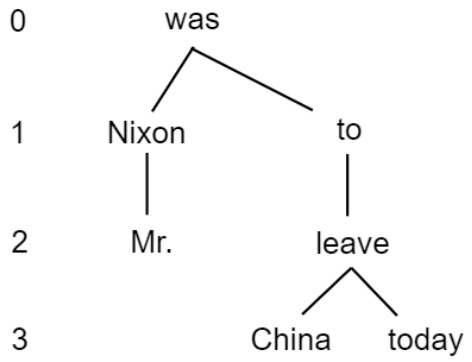

Figure 29 Arbre de dépendance (exemple de Jing et Liu (2015)) avec les hauteurs

La distance hiérarchique est une métrique équivalente à la profondeur de Köhler et Altmann (2000). Pourtant, leurs valeurs ne sont pas identiques à cause des analyses structurelles en amont qui sont différentes. Les analyses de Köhler et Altmann sont fondées sur l'approche de constituants et celles de Jing et Liu (2015) sur l'approche de dépendances. Les exemples (23) et (24) montrent les différentes analyses pour une structure du groupe nominal, et les valeurs de la profondeur ou de la distance hiérarchique. Dans (23), le constituant NP est composé de N(nom) et D (déterminant), ainsi, la profondeur de D et la profondeur de N sont de 1 (si nous considérons que le nœud NP est la racine). Cependant, cette structure en dépendance est analysée seulement comme une relation de dépendance dont le gouverneur est N et le dépendant est D : N→D. Si N est considéré comme la racine, la distance hiérarchique de N est de 0 et elle est de 1 pour D.

(23) Analyse en constituants de Köhler et Altmann dans un groupe nominal (NP) :

NP =N+D

Profondeur de D = 1

Profondeur de N = 1

(24) Analyse en dépendance de Jing et Liu (2015) :

N→D

Longueur hiérarchique D = 1

Longueur hiérarchique N = 0

La valeur de distance hiérarchique dépend aussi du schéma d'annotation (section 2.4.1). La figure 29 montre l'arbre de dépendance de Jing et Liu (2015) et ses valeurs de distance hiérarchique. Les analyses de cet arbre sont similaires au schéma SUD (section 2.4.1), qui prend les mots fonctionnels comme tête syntaxique. Par exemple, entre *was* et *leave*, il s'agit de deux relations (*was*→*to*, *to*→*leave*), ainsi la distance hiérarchique de *leave* est 2. Cependant, les distances hiérarchiques seront différentes dans les corpus annotés en schéma UD, car ce dernier prend les mots pleins comme têtes syntaxiques. Ainsi, la figure 30 montre la même phrase analysée en UD. Nous pouvons observer que la hauteur de l'arbre est réduite à 2. Les distances hiérarchiques sont différentes par rapport à la figure 30 : il s'agit d'une réduction des valeurs pour les mots pleins (*leave, china, today*) et d'une augmentation des valeurs pour le mot fonctionnel (*to*).

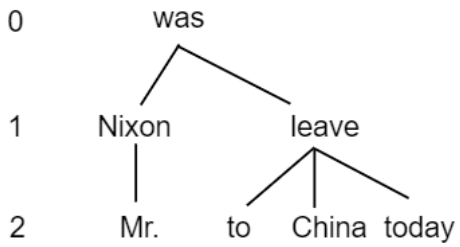

Figure 30 Analyse du type UD

Jing et Liu (2015) ont calculé la distance hiérarchique moyenne (MHD) et la distance de dépendance moyenne (MDD) (voir section 3.3.3.2) dans le corpus *Prague Czech-English Dependency Treebank* 2.0 (PCEDT, 2.0) en anglais et tchèque. Les résultats montrent que MDD et MHD sont inférieures à 4 dans les deux langues. Ils ont également constaté qu'elles ont des comportements différents pour les deux langues :

> *"English tends to reduce the syntactic complexity in the hierarchical dimension, whereas Czech prefers to lessen the processing load in the linear dimension."*

D'ailleurs, selon Jing et Liu (2015), la distance de dépendance moyenne (MDD) et la distance hiérarchique moyenne (MHD) se fondent sur des interprétations cognitives distinctes, et MHD mesure plutôt la complexité basée sur la production :

> *"To be more specific. MDD is actually a comprehension-oriented metric that measures the difficulty of transforming linear sequences into layered trees, whereas MHD is a production-oriented metric that measures the complexity of transforming hierarchical structures to string of words."*

3.3.2.4 Commentaires

Au niveau formel, nous avons principalement présenté deux types de métriques pour étudier la

complexité structurelle d'un arbre. La première consiste à observer verticalement sa hiérarchie, comme la profondeur de Yngve (1960), la profondeur de Köhler et Altmann (2000) et la distance hiérarchique (Jing & Liu, 2015). L'autre consiste à observer le nombre d'enfants pour chaque nœud, comme l'arité (Courtin & Yan, 2019) et la complexité syntaxique (Köhler & Altmann, 2000). Certaines métriques sont formellement équivalentes. La profondeur de Köhler et Altmann est équivalente à la distance hiérarchique de Jing et Liu, mais la première est basée sur l'arbre de constituants plat et la seconde sur l'arbre de dépendance. L'arité sur l'arbre de dépendance et la complexité syntaxique sur l'arbre de constituants plat sont également équivalents par définition. Les observations des travaux précédents ont pour point commun de constater une limitation très probablement causée par des contraintes mémorielles.

Les travaux de cette thèse se concentrent sur les études de la complexité syntaxique dans le cadre du traitement des phrases. Selon Abney (1989), le traitement de la phrase chez l'humain est progressif, de gauche à droite. Ainsi, il est important que les métriques de la complexité syntaxique contiennent des informations linéaires. Les métriques pour l'arbre ordonné prennent en compte l'ordre linéaire, elles nous permettent de voir la complexité syntaxique pour chaque moment de traitement. Et donc dans la section 3.3.3, nous allons aborder les métriques pour les arbres ordonnés.

3.3.3 Métriques pour arbre ordonné

Nous rappelons que la structure arborescente utilisée dans la section précédente ne contient pas d'informations sur la mise en place de chaque élément syntaxique dans la séquence linéaire.

La figure 31 montre un arbre de dépendance, il ne contient que de l'information de structure hiérarchique. Nous pouvons linéariser les mots de la structure dans n'importe quel ordre :

a. Poursuit chat le souris la.
b. Chat souris le poursuit la.
c. Le chat poursuit la souris.
d. La souris poursuit le chat.
e. Autres possibilités en combinatoire

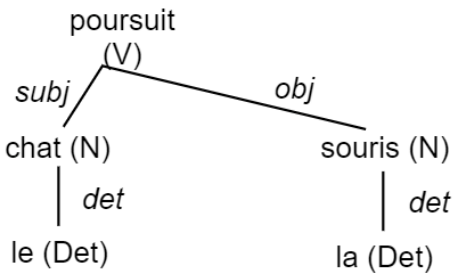

Figure 31 Arbre de dépendance non-ordonné

Nous savons que l'ordre acceptable de la phrase est celui en *c*, ce qui signifie qu'une structure hiérarchique n'est pas transformée de manière aléatoire en une séquence, mais est contrainte de quelque manière. Les différents choix des linéarisations[31] d'éléments syntaxiques affecteront la complexité syntaxique de la phrase. Seules les phrases dont les éléments syntaxiques sont linéarisés dans certains ordres peuvent être traitées avec succès et finalement comprises. Ainsi, il est également important de connaître la position linéaire de chaque élément syntaxique dans la séquence.

Plus précisément, pour les arbres de constituants, nous pouvons diviser la séquence linéaire avec des crochets afin de montrer les constituants et leur hiérarchisation. Par exemple, S [NP [Le chat] VP [poursuit NP [la souris]]] est une représentation linéarisée d'un arbre de constituants binaires[32]. La métrique PCD, décrite en 3.3.3.1, est proposée à partir de tels arbres ordonnés. Concernant l'arbre de dépendance, on peut associer la structure syntaxique à la structure linéaire[33] (voir aussi Kahane & Gerdes, 2021, section 3.5.9). Cela nous permet d'obtenir l'arbre de dépendance ordonné, comme le montre la figure 32. Les métriques en 3.3.3.2 et 3.3.3.3 utilisent les arbres de dépendance ordonnés.

---

[31] Selon Kahane et Gerdes (2021) : *« La linéarisation est modélisée comme une correspondance entre deux structures de niveaux différents : une structure hiérarchique non ordonnée — la structure syntaxique — et une structure ordonnée. »*

[32] Ce type de représentation a ses limites, notamment lorsqu'il s'agit des phrases non-projectives (voir la notion de non-projectivité dans la section 3.4.1).

[33] On peut lire une séquence linéaire en tant que structure linéaire, c'est-à-dire que chaque nœud/mot a une relation de précédence avec le nœud qui le suit (Kahane & Gerdes, 2021, section 3.5.9)

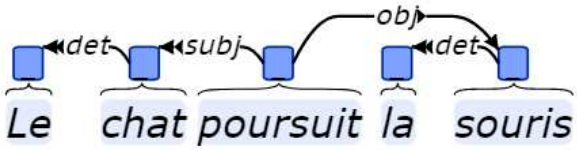

Figure 32 Arbre de dépendance ordonné

3.3.3.1 *Phrasal combination domain*

Hawkins (1994) a proposé de mesurer le *phrasal combination domain* (PCD) dans les arbres de constituants afin d'étudier l'efficacité et la complexité de la grammaire. Comme l'étude de la complexité syntaxique, celle de l'efficacité et la complexité de la grammaire s'appuie sur la performance linguistique (voir 2.2 pour la notion de performance linguistique). Ce qui est différent avec l'étude de la complexité syntaxique est qu'elle consiste plutôt à étudier le lien entre la grammaire et la performance linguistique. Plus précisément, elle se concentre sur l'étude de la préférence de l'ordre des mots pour un certain type de structure syntaxique, et sur les hypothèses entre la performance et la grammaire.

Le ***Phrasal combination domain*** pour un nœud mère (*M*) et ses constituants immédiats (*ICs*) consiste en plus petite chaîne des séquences, qui permettent à l'analyseur syntaxique de construire le nœud mère (*M*) et ses constituants immédiats (*ICs*).

L'exemple (25) montre deux linéarisations d'une même structure syntaxique. La linéarisation du groupe verbal dans *a* est une construction *short before long* (SL) (voir aussi 1.3.2). Dans ce cas, pour réussir à construire le nœud de VP avec deux sous constituants PP1 et PP2, il faut que l'analyseur traite au moins cinq mots (de *looked* jusqu'à *into* qui est la tête de PP2). Ainsi le *PCD* sera de 5. Quant au cas de *b* qui contient une construction *long before short* (LS), le *PCD* sera de 9 (de *looked* jusqu'à *through* inclu).

(25) (a). **SL:** *The gamekeeper* **VP**[*looked* **PP1**[*through his binoculars*] **PP2**[*into*

1 2 3 4 5

the blue but slightly overcast sky]].

(b). **LS:** *The gamekeeper* **VP**[*looked* **PP2**[*into the blue but slightly overcast*

1 2 3 4 5 6 7

sky] **PP1**[*through his binoculars*]].

8 9

À partir de cela, l'efficacité syntaxique du traitement de ces deux linéarisations peut être calculé en ratio *IC-to-Word* : la valeur de *PCD* divisée par son nombre de constituants immédiats. Pour (a), le ratio *IC-to-Word* est 3/5 (60%) et pour (b) est 3/9 (33%).

Hawkins a fait des études sur la distribution des cas d'ordre de l'exemple 3.3 dans un corpus anglais. Il a trouvé dans le corpus seulement 15% du cas *LS*, contre 82% du cas *SL*.

Des études similaires ont été faites dans les données de langues à tête finale comme le japonais. Selon Yamashita et Chang (2001) et Hawkins (1994, 2004), comme le verbe se trouve à droite des deux groupes prépositionnels, le cas du *LS* est largement préféré à *SL* en japonais. Lorsque le verbe est à droite, l'efficacité du traitement est supérieure dans l'ordre *LS* à celui du cas *SL*.

Minimisation de PCD

Selon Hawkins (1994), dans les *PCDs* la distribution des valeurs montre une plus grande fréquence pour les valeurs faibles dans les langues naturelles (le principe de *Early Immediate Constituents*) :

> *"EARLY IMMEDIATE CONSTITUENTS (EIC)*[34]
> *The human processor prefers linear orders that minimize PCDs (by maximizing their IC-to-word ratios), in proportion to the minimization difference between competing orders. (Hawkins, 2004)*

Cette hypothèse est similaire au phénomène de *heavy-np shift*. Celui-ci est proposé premièrement par les encyclopédistes Buffier (1709) et Bauzée (1765). Bauzée l'a expliqué en

[34] Le principe EIC est généralisé de Hawkins (2004) en principe de *Minimize Domaine* (MiD)

mettant en cause la compétition entre deux constituants :

> *« De plusieurs compléments qui tombent sur le même mot, il faut mettre le plus court le premier après le mot complété ; ensuite le plus court de ceux qui restent, & ainsi de suite jusqu'au plus long de tous qui doit être le dernier. […] il importe à la netteté de l'expression, cujus summa laus perspicuitas, de n'éloigner d'un mot, que le moins qu'il est possible, ce qui lui sert de complément. Cependant quand plusieurs compléments concourent à la détermination d'un même terme, ils ne peuvent pas tous le suivre immédiatement ; & il ne reste plus qu'à en rapprocher le plus qu'il est possible celui qu'on est forcé d'en tenir éloigné : c'est ce que l'on fait en mettant d'abord le premier celui qui a le plus de brièveté, & réservant pour la fin celui qui a le plus d'étendue. »*

Les résultats dans les différents types de langues ont prouvé que la métrique *PCD* permet de soutenir le cadre théorique de la *Performance-Grammar Correspondance Hypothesis* (Hawkins, 2004) qui considère que la performance (l'efficacité du traitement) est un facteur qui influence l'ordre des mots (la grammaire). L'efficacité du traitement est différente selon les deux (ou plusieurs) linéarisations possibles d'une structure, on a tendance à choisir la linéarisation où l'efficacité est la plus grande.

De plus, en se basant sur des observations dans différentes langues, Hawkins (2004) a proposé une hypothèse du traitement des phrases qui est proche de celle de la localité de dépendance (Gibson, 2000) et de la minimisation de la longueur de dépendance (voir la section suivante) (Liu, 2008 ; Gildea & Temperley, 2010 ; Futrell et al, 2015 ; Ferrer-i-Cancho, 2006 ;2017) :

> *"More generally, we can hypothesize that the processing of all syntactic and semantic relations prefers minimal domains."*

Pourtant, la métrique *PCD* ne permet pas vraiment de mesurer une complexité pour une phrase. Elle est souvent utilisée pour déterminer une préférence dans l'ordre des mots. De plus, les travaux utilisant cette métrique se pratiquent en comparant deux ou plusieurs linéarisations, ce

qui ne permet pas d'aboutir à vérifier la limitation mémorielle dans le traitement des phrases.

3.3.3.2 Longueur de dépendance

La longueur de dépendance (ou distance de dépendance) est le nombre de mots entre le gouverneur et son dépendant dans une relation de dépendance. (Hudson, 1995; Liu, 2008 ; Ferrer-i-Cancho, 2006 ; Gildea & Temperley, 2010 ; Futrell et al, 2015).

La figure 33 montre un arbre de dépendance dans un treebank du projet UD, et la longueur de dépendance pour chaque mot.

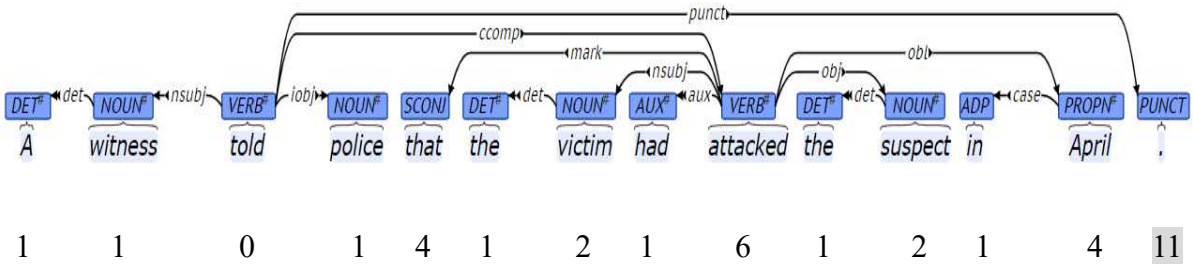

Figure 33 Arbre de dépendance avec les longueurs de dépendance

L'une des premières études sur la longueur de dépendance est celle de Hudson (1995), où elle est proposée comme l'une des métriques pour analyser la complexité syntaxique. Hudson (1995) a mis en relation entre la longueur de dépendance et le dégré de la dégradation dans la mémoire pour un mot déjà traité (voir aussi, Gibson, 2000) :

> *"A realistic model of a human parser would not assume that the whole sentence is preserved, being shunted between the stack and the current buffer, right to the end. ...Incremental parsing extracts meaning as quickly as possible, after which the words themselves are of no value and can be abandoned. However 'forgetting' is clearly not a clean deletion operation, but a gradual deterioration; so words get harder and harder to retrieve, without necessarily being lost altogether. It is the words on the stack that are*

deteriorating in this way, so being on the stack does not guarantee equal accessibility."

Dans cette section, nous allons présenter les travaux de longueur de dépendance en deux parties : les études quantitatives de la longueur de dépendance moyenne de Liu (2008), ainsi que celles de Gildea et Temperley (2010) et de Futrell et al. (2015) qui s'interrogent davantage sur le lien entre la longueur de dépendance et l'ordre des mots.

Il faut aussi noter que les différentes analyses structurelles pourraient mener à des valeurs différentes pour la longueur de dépendance. Par exemple, pour les treebanks d'UD utilisés par Futrell et al (2015) et les treebanks utilisés par Liu (2008), la préposition est considérée comme le dépendant dans une relation *case*, comme dans la figure 33 : {in <-*case*- April}. La tête du mot *April* est le verbe *attacked*, ainsi, la longueur de dépendance est de 4 pour le mot *April*. Pourtant dans les treebanks utilisées par Gildea et Templerly (2010) et Ferrer-i-Cancho (2006), la préposition est considérée comme la tête : {in -*case*-> April}, la longueur de dépendance est de 1.

Longueur de dépendance : les études quantitatives

Selon Liu (2008), on peut avoir la longueur de dépendance moyenne (aussi dite distance de dépendance moyenne, angl : *sentence mean dependency distance*, $MDD_{sentence}$) de cette phrase par les formules F2 et F3. Dans la formule F2, DD_i est la longueur signée (positive si le gouverneur est après son dépendant, négative sinon) pour la dépendance *i*, *n* est la longueur de la phrase. Ainsi, le nombre de relation est *n-1* comme la racine n'a pas de gouverneur. Il est également possible de calculer la longueur de dépendance moyenne pour un texte, exprimée par la formule F3, où *n* est la longueur du texte (le nombre de mots dans le texte), *S* est le nombre de phrases, ainsi, le nombre de dépendances total dans le texte est $n - S$.

F2. $MDD_{sentence} = \frac{1}{n-1}\sum_{i=1}^{n}|DD_i|$

F3. $MDD_{text} = \frac{1}{n-S}\sum_{i=1}^{n}|DD_i|$

Avec ces deux formules, si l'on ne compte pas les relations de ponctuation, dans l'exemple de la figure 33, le MDD est de 2,17, et la longueur maximale est de 6 pour le verbe *attacked*.

Liu (2008) a souligné que la longueur maximale est théoriquement plus pertinente si l'on veut étudier la limitation de la longueur de dépendance. Mais elle n'est pas suffisamment stable entre les différentes langues. Selon Liu (2008), l'avantage de MDD est qu'il permet de capturer la difficulté moyenne considérant le traitement de chaque mot :

> *"Hudson (1995) proposes a simple way to measure the load due to open dependencies is to weight each open dependency equally* (voir aussi section 3.3.3.3 et 4.3.4), *and to sum the open dependencies after each word (for example, I actually live in Beijing)… we can create a series of dependency density 1+2+1+1+0, whose companion of dependency distance is 2+1+0+1 +1. Thus, MDD and the mean number of items that are kept in Headlist (working memory) during the parsing process are positively correlated. Using MDD, we can measure relative difficulty of a sentence. Sometimes it is probably not more precise than using maximum DD to measure absolute difficulty to build a dependency link, but it works with corpus and real text."*

Liu (2008) a calculé la longueur de dépendance moyenne dans 20 langues[35]. En plus des treebanks originaux, Liu a généré deux versions de treebanks artificiels comme modèle de base.

Liu dit d'un modèle de base *RL1* qui contient des treebanks générés aléatoirement :

> *"…within each sentence we select one word as the root, and then for every other word we randomly select another word in the same sentence as its governor, disregarding syntax and meaning."*

Liu a constaté que la longueur de dépendance moyenne dans 20 treebanks est entre 1,98 (et c'est pour le roumain) et 3,662 (qu'on observe pour le chinois), 3,356 (pour le turc) et 18,474

[35] Les données de treebanks sont venues de *CoNLL-X Shared Task on Multilingual Dependency Parsing* 6 (2006, 2007). Les treebanks sont convertis en un même schéma d'annotation. (Liu, 2008)

(obtenu pour l'arabe) pour les treebanks générés de manière aléatoire (*RL1*).

Un autre modèle de base *RL2* contient des treebanks générés aléatoirement en tenant compte de la contrainte de projectivité (voir aussi 3.4 pour la question de la projectivité). Les expériences de Liu montrent que la longueur de dépendance moyenne de *RL2* est inférieure à celle de *RL1*, mais supérieure à celle des treebanks originaux.

Pour conclure, les résultats de Liu (2008) ont montré que la longueur de dépendance tend à être minimisée dans les langues naturelles, et que la contrainte formelle de la projectivité entraîne des longueurs de dépendance plus petites.

Minimisation de longueur de dépendance et l'ordre des mots

Temperley (2008) et Gildea et Temperley (2010) suggèrent que la longueur de la dépendance est un facteur qui influence l'ordre des mots. Nous reprenons l'exemple (25) de la section 3.3.3.1. Par rapport à ce que Hawkins (2004) a proposé par le principe *EIC* (voir 3.3.3.1), Gildea et Temperley pensent que la longueur de dépendance dans la phrase *SL* (Figure 34) est inférieure à *LS* (Figure 35) (26 contre 33), donc la configuration *SL* est préférée.

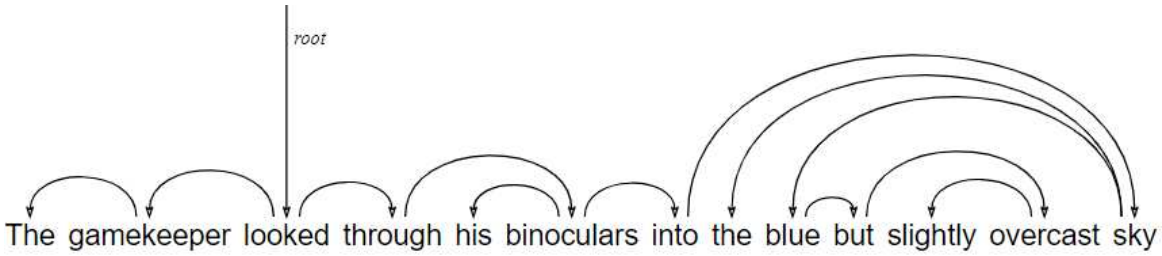

Figure 34 Longueur de dépendance de la phrase : 1+1+1+2+1+1+6+5+4+1+2+1 = 26

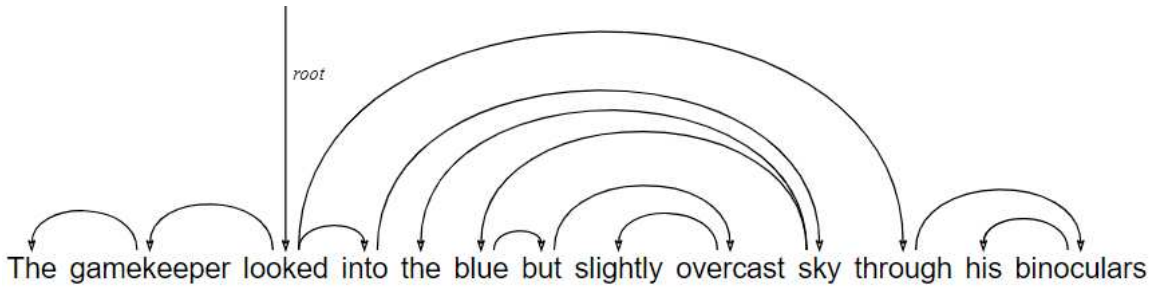

Figure 35 Longueur de dépendance de la phrase : 1+1+1+8+6+5+4+1+2+1+2+1 = 33

Pour étudier l'impact de la minimisation de la longueur de dépendance sur l'ordre des mots, Gildea et Temperly (2010) ont comparé la longueur de la dépendance dans trois types de données : les arbres linéarisés générés par un algorithme qui optimise la longueur de la dépendance, les arbres linéarisés aléatoirement et les arbres avec l'ordre original du corpus (*Wall Street Journal* dans *Penn Treebank*).

L'objectif de Gildea et Temperley (2010) est d'étudier si les arbres ordonnés originaux ont des longueurs de dépendance les plus courtes. Les résultats montrent que la longueur de dépendance pour chaque phrase dans le corpus original était en moyenne de 47,5, contre 82,7 pour la version à ordre des mots aléatoire. Cependant, les longueurs de dépendance des corpus d'origine ont dépassé celles des arbres en ordre optimal, qui elles ont un score en moyenne de 33,5.

Le second travail de Gildea & Temperley (2010) consiste à ajouter les règles grammaticales dans l'algorithme optimal pour déterminer l'ordre des mots optimal sur les arbres[36] :

> *"... each head-dependent set (a syntactic head type and a set of dependent types) to be ordered in the same way on each occurrence."*

Ce modèle a obtenu une longueur de dépendance par phrase de 42,5, ce qui est beaucoup plus proche de celle du corpus réel (47,5). Cela montre que la grammaire joue également un rôle pour l'ordre des mots, qui est prioritaire par rapport à la minimisation de la longueur de

[36] Les informations d'ordre que cet algorithme utilise proviennent du corpus original.

dépendance.

Futrell et al. (2015) ont également comparé la longueur de dépendance dans les treebanks originaux avec les treebanks générés pour chaque longueur de phrase[37]. Les résultats de Futrell et al. (2015) ont mené à une conclusion similaire à celles de Liu (2008) et de Gildea et Temperley (2010). Ils pensent que la minimisation de la longueur de dépendance est un phénomène universel, et que les humains préfèrent les ordres de mots qui minimisent la longueur de dépendance, tout en respectant la grammaire de base.

3.3.3.3 Densité de dépendance (*angl :Dependency density*)

Hudson (1995) a proposé une métrique qui s'appelle densité de dépendance, qui est le nombre de dépendances « ouvertes » à un moment donné. Dans les chapitres 4 et 5, nous utilisons le terme *flux requis* qui en sera un équivalent théorique.

Hudson (1995) pense qu'une dépendance incomplète est plus difficile à traiter qu'une dépendance déjà complète. Donc, la difficulté au moment du traitement d'un mot dépend du nombre de dépendances incomplètes :

> *"It seems reasonable to assume that dependencies place a greater load on working memory when they are still open than once they are closed, because*

[37] Futrell et al. (2015) ont utilisé quatre types de données dans 37 langues :

- Les arbres originaux de 37 treebanks d'UD
- Les arbres linéarisés en optimisant la minimisation de longueur de dépendance (angl : *optimal baseline*) (Gildea & Temperley, 2010)
- Les arbres linéarisés aléatoirement (angl : *Free word order base-line*) (Gildea & Temperley, 2010)
- Les arbres linéarisés aléatoirement mais avec ordres des mots fixes (angl : *fixed word order baseline*). Dans les travaux de Furtrell et al. (2015), un poids [-1,1] est attribué aléatoirement à chaque type de relation, ensuite l'algorithme linéarise la structure : *"starting at the root node, collect the head word and its dependents and order them by their weight, with the head receiving weight 0. The repeat the process for each dependent, keeping the same weights. This procedure creates consistency in word order with respect to relation types."*

a closed dependency has achieved its main function of guiding the hearer to a semantic structure. Once the semantic structure is in place, and especially when it has been pragmatically exploited, the syntactic dependency can be allowed to disappear from working memory."

La Figure 36 est un exemple anglais analysé en dépendance de Hudson (1995). Il l'a utilisé pour montrer les valeurs de densité de dépendance pour chaque mot. Le nombre de dépendances incomplètes contient les dépendances qui attendent encore leur gouverneur, et les dépendances qui attendent un dépendant obligatoire.

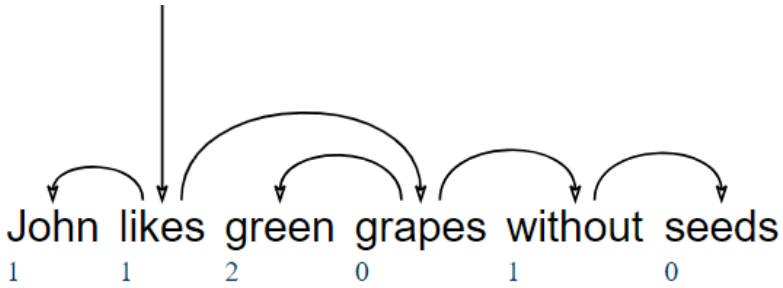

Figure 36 Exemple avec les valeurs de densité de dépendance, utilisé par Hudson (1995)

Par exemple, pour le mot *John*, comme il attend encore un verbe pour établir une dépendance de type sujet, ainsi, la valeur de densité de dépendance est de 1. Pour le mot *likes*, il attend un dépendant de type objet qui est obligatoire, ainsi la valeur de densité de dépendance est de 1. Pour le mot *grapes*, on observe que la dépendance {grapes ->without} est une relation adjointe, qui n'est pas une dépendance obligatoire, ainsi, la valeur de densité de dépendance est de 0.

Selon Hudson (1995), les dépendances incomplètes doivent être stockées temporairement dans la mémoire de travail. Comme la difficulté du traitement est due au stockage mémoriel, plus les dépendances sont stockées, plus il est difficile de les traiter. Dans 4.3.4, nous reparlerons de cette métrique et de la façon de la calculer dans les treebanks. Ensuite, nous allons présenter les distributions des valeurs de cette métrique dans les treebanks dans 5.5.3.

3.4 Contraintes formelles

Les contraintes formelles sont un sujet important dans le domaine de la complexité syntaxique. Cette section ne concerne pas une étude spécifiquement des métriques, mais nous nous focalisons ici sur le lien entre la complexité syntaxique et les contraintes formelles. Précisément, nous allons commencer par présenter la contrainte de projectivité (section 3.4.1) et la contrainte *bien-nichée* (angl : *well-nested*) (section 3.4.2). Ensuite, la section 3.4.3 se focalisera elle sur l'interprétation de la minimisation de la longueur de dépendance en relation à la projectivité.

3.4.1 Projectivité de l'arbre de dépendance

En observant la manière dont les structures de dépendance se linéarisent, Ihm & Lecerf (1960) ont proposé la projectivité comme une contrainte formelle dans les langues naturelles.

La projectivité a été étudiée par de nombreux auteurs (Ihm & Lecerf, 1960 ; Marcus, 1965 ; Melʹčuk 1988 ; Kahane, 2001 ; Kuhlmann & Nivre, 2006 ; Havelka, 2007), et nous utilisons ici les définitions générales de Ihm & Lecerf (1960) et de Kuhman & Nivre (2006) :

Projection : *i* est un nœud dans l'ensemble des nœuds *V* de l'arbre de dépendance, l'ensemble des nœuds linéarisés dominés par *i* est la projection de *i*.

Projection continue : si $y \in V$ et *y* se trouve dans le domaine de la projection de *i*, *y* doit faire partie de **sous-arbre**$_i$. Pour maintenir la projection continue pour le nœud *i* dans sa linéarisation, il faut mettre tous les nœuds du **sous-arbre**$_i$ autour de *i*.

Projectivité : Un graphe de dépendance est projectif, si les projections de ses nœuds sont continues.

S'il s'agit d'une projection non continue pour un nœud dans un arbre, cet arbre est non projectif. La figure 37 montre un arbre, avec sa linéarisation non projective (Figure 38) représentée en *A′*, *B′*, *C′* et *D′*. On peut constater que la projection du nœud *C* n'est pas continue : le domaine

du *C'* est entre *A'* et *C'*, pourtant *B'* étant entre *A'* et *C'* ne fait pas partie dans **sous-arbre**$_{C'}$.

Visuellement, la non-projectivité peut être observée lorsque deux arcs de dépendance se croisent. Dans la figure 38 , il s'agit de l'arbre de dépendance ordonné pour la figure 37, deux croisements sont observés : {*B'* -> *D'*} et {*A'* <-*C'*}, {*A'* <-*C'*} et {*Racine* -> *B'*}

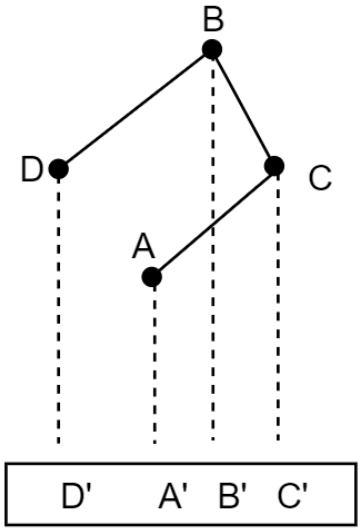

Figure 37 Arbre de dépendance avec sa liénarisation non projective

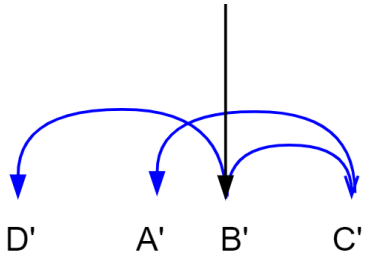

Figure 38 Arbre de dépendance ordonné

Les structures non projectives ne sont pas totalement interdites. Nous pouvons trouver dans le tableau 5 des exemples de constructions en français qui pourraient avoir la propriété non projective (Kahane & Gerdes, 2021).

Types	Exemples
Montée des clitiques	*Zoé **lui** a parlé. / Zoé **en** connaît la moitié.*
Comparatif et superlatif	*un plus beau livre **que le mien***
Extraction	***A quel endroit** Zoé a-t-elle envie d'habiter ?*

Tableau 5 Constructions non projectives (Kahane & Gerdes, 2021)

Kuhman et Nivre (2006) ont mené des études quantitatives sur la non-projectivité dans le treebank *Prague Dependency Treebank* (PDT) (Hajic et al., 2001) et le treebank *Danish Dependency Treebank* (DDT) (Kromann, 2003). Les résultats montrent que 15 % des arbres sont non-projectifs dans le DDT et 23 % dans le PDT.

3.4.2 Non projectivité et contrainte bien-nichée

Différentes linéarisations pourraient créer des cas non projectifs distincts, mais ces cas n'ont pas la même complexité syntaxique. Afin d'analyser la non-projectivité, il est possible d'observer le nombre de croisements des dépendances. Nous considérons que la phrase est plus difficile à traiter si elle est constituée de plusieurs relations croisées. Ainsi, l'exemple dans la figure 40 est considéré comme plus difficile à traiter que celui de la figure 39 car le nombre de croisements dans la phrase de la figure 39 est de 1, contre 3 pour la figure 40.

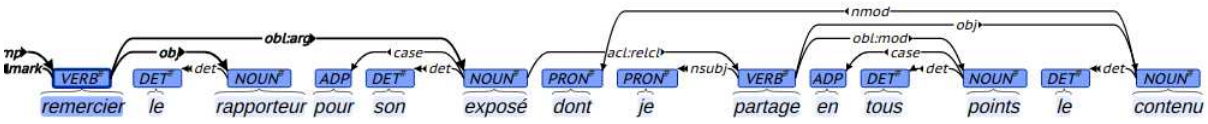

Figure 39 Exemple (UD_French-Sequoia, sent_id = Europar.550_00005)

C(nmod, acl:relcl) : croisement entre {dont <-*nmod*- contenu} et {exposé -*acl:relcl*-> partage}

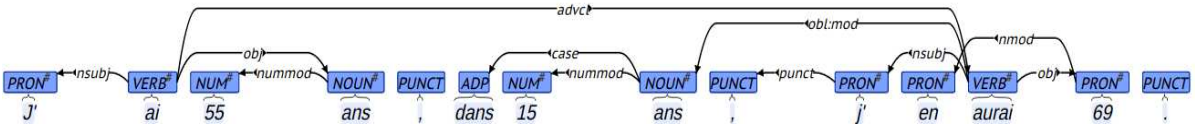

Figure 40 Exemple (UD_French-Sequoia, sent_id = Europar.550_00488)

C(nmod, advl) : {en <-*nmod*- 69} et {ai -*advcl*->aurai}

C(nmod, obl:mod) :{ en <-nmod- 69} et {ans <-*obl:mod*- aurai}

C(nmod, nsbuj) : { en <-nmod- 69} et {j' <-*nsubj*- aurai}

Une autre façon d'analyser la non-projectivité est d'observer son type du croisement. On peut constater que les deux dépendances de l'exemple dans la figure 39 font partie d'un même sous-arbre (**sous-arbre**$_{\text{exposé}}$). Mais concernant l'exemple de la figure 40, on peut constater que les dépendances {en <-*nmod*- 69} et {ai -*advcl*->aurai} sont dans **sous-arbre**$_{\text{ai}}$. Pourtant ce n'est pas la même situation pour les dépendances {en <-*nmod*- 69} et {ans <-*obl:mod*- aurai}, ainsi que pour les dépendances {en <-*nmod*- 69} et {j' <-*nsubj*- aurai}. Dans ces derniers cas, les deux dépendances ne sont pas dans un même sous-arbre, et pourtant elles partagent un même nœud supérieur *aurai*.

On pourrait classer les différents types de croisement par leur niveau de non-projectivité. Dans la figure 41, il s'agit de trois types de croisement pour un arbre non ordonné (les segments rouges sont des croisements).

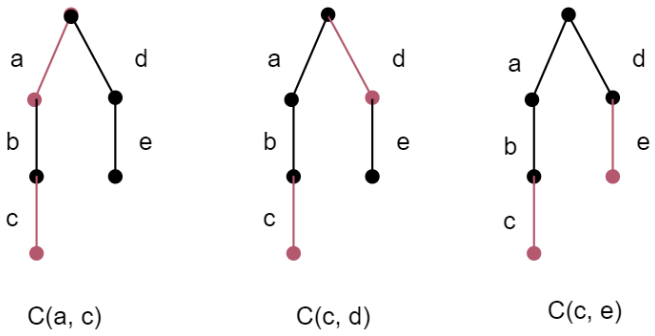

Figure 41 Trois types de croisement pour un arbre non ordonné

Pour le croisement de la dépendance *a* et la dépendance *c, C(a, c),* nous pouvons constater qu'il s'agit du dépendant d'une relation (*a*) qui fait partie d'un sous-arbre du dépendant de l'autre relation (*c*). Nous considérerons ici que le niveau de non-projectivité est de 1. L'exemple de la

figure 39, *C(nmod, acl: relcl)* relève également de ce type de croisement.

En ce qui concerne le cas pour *C(c,d)*, le dépendant de *c* n'est pas dans le sous-arbre du dépendant de *d*, mais il fait partie du sous arbre du gouverneur de *d*. Alors le niveau de non-projectivité est de 2. Dans l'exemple de la figure 40, *C(nmod, obl:mod)* et *C(nmod, nsubj)* sont au niveau 2, et *C*(*nmod, advcl*) est au niveau 1.

Le cas du *C*(*c,e*) est encore différent par rapport aux situations de niveau 1 et de niveau 2. En effet *c* et *e* sont dans deux sous-arbres disjoints, et ce type de croisement est au niveau 3.

Le cas du niveau 1 est le plus facile à traiter, car les quatre nœuds sont dans un même constituant commun. Concernant les cas du niveau 2 et 3, il faut traiter deux constituants. De plus, le cas du niveau 3 est de fait le plus complexe à traiter par rapport au deux autres niveaux, car les deux constituants sont disjoints.

Une construction relevant du cas de niveau 3 est mal-nichée (angl : *ill-nested*[38]). Si l'arbre ne contient pas de cas *mal-niché*, il est dit *bien-niché*. La figure 42 montre l'arbre linéarisé avec le croisement de *c* et *e*, ce qui est un exemple de structure mal-nichée.

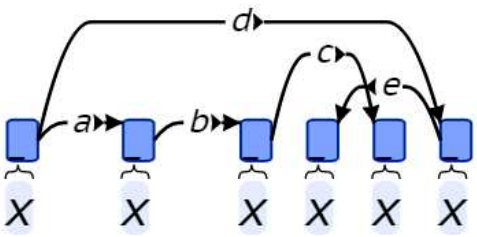

Figure 42 Non-projectivité à niveau 3, structure mal-nichée

La contrainte bien-nichée proposée par Bodirsky et al. (2009), Kuhman et Nivre (2006) et Havelka (2007) a été étudiée dans les langues. Selon les résultats de Kuhman et Nivre (2006), la contrainte bien-nichée peut inclure 99.98% d'arbres dans DDT et PDT, mais cela ne

[38] Définition de mal-niché (angl : *ill-nested*) (Bodirsky et al., 2009; Havelka, 2007): *"A dependency tree T = (V,→,≤) is ill-nested if there are disjoint subtrees T1, T2 of T and nodes x1, y1 ∈ T1 and x2, y2 ∈ T2 such that x1< x2 < y1< y2. Otherwise the dependency tree T is well-nested."*

correspond qu'à 85% (DDT) et 77% (PDT) de cas de projectivité. Ainsi on peut dire que la non-projectivité n'est pas totalement interdite, mais que les constructions mal-nichées sont rares.

3.4.3 Projectivité et minimisation de longueur de dépendance

Dans cette section, nous nous focalisons sur les travaux de Gómez-Rodríguez et Ferrer-i-Cancho (2017) qui étudient la rareté de la non-projectivité dans le langage naturel.

Gómez-Rodríguez et Ferrer-i-Cancho (2017) s'intéressent à la relation entre la longueur de dépendance et le nombre de croisements. Leur hypothèse est que là où les longueurs de dépendance tendent à être minimisées, les croisements sont plus rares.

Pour vérifier cette hypothèse, deux modèles de prédiction du croisement sont proposés. Le premier est un modèle basique (angl : *base-line*), exprimé dans l'équation F4. Q représente l'ensemble de toutes les paires de dépendances. $C(e_1, e_2) = 1$ signifie que les deux dépendances e_1 et e_2 se croisent. $C(e_1, e_2) = 0$ signifie que e_1 et e_2 ne se croisent pas. Le modèle permet de calculer l'espérance pour le nombre de croisements $E_0[C]$, en fonction de la probabilité du croisement pour chaque paire de relations $p(C(e_1, e_2) = 1)$.

F4. Modèle de base :

$$E_0 = \sum_{(e_1,e_2)\in Q} E[C(e_1, e_2)]$$

$$= \sum_{(e_1,e_2)\in Q} p(C(e_1, e_2) = 1)$$

Le second modèle prend en compte les longueurs de dépendance de chaque paire de relations croisées, comme l'équation F5. $d(e_1)$ est la longueur de dépendance de e_1 et $d(e_2)$ pour la longueur de dépendance de e_2. L'espérance pour le nombre des croisements $E_2[C]$ dépend

de la probabilité conditionnelle sachant les deux longueurs de dépendance : $p(C(e_1, e_2) = 1 \,|d(e_1), d(e_2))$.

F5. Modèle considérant la longueur de dépendance :

$$E_2 = \sum_{(e_1,e_2)\in Q} p(C(e_1, e_2) = 1 \,|d(e_1), d(e_2))$$

Les erreurs de prédiction des E_0 et de E_2 sont Δ_0 et Δ_2. Elles peuvent être obtenues en comparant le nombre de croisements prédits des deux modèles avec le nombre de croisements dans le corpus réel, Treebank (version 2.0) de HamleDT (Zeman et al., 2014 ; Rosa et al., 2014) en 30 langues.

Selon Gómez-Rodríguez et Ferrer-i-Cancho (2017), le modèle *base-line* (E_0) surestime le nombre de croisements lorsque la longueur de phrase dépasse 6. L'erreur du modèle considérant la longueur de dépendance (E_2) Δ_2 est bien inférieure (5% contre 33% pour Δ_0) à celle du modèle de base. Le modèle E_2 est plus performant par rapport à E_0, parce qu'on observe que Δ_2 est 6 fois moindre que Δ_0. Ces résultats prouvent que prendre en compte des longueurs de dépendance permet de mieux prédire le nombre de croisements dans les langues. D'après la section 3.3.3.2, on sait que la longueur de dépendance tend à être minimisée, ainsi, la rareté de la non-projectivité serait un effet de la minimisation de dépendance.

4 Flux de dépendance

4.1 Introduction

Dans ce chapitre, nous verrons que le flux de dépendance (voir aussi 0.2.3 dans l'introduction générale de cette thèse) nous permet d'élaborer des métriques sur la complexité syntaxique correspondant à différentes hypothèses cognitives. Tout d'abord, nous allons présenter la représentation du flux sous forme de matrice (Botalla, 2014) dans la section 4.2. Cette représentation nous aidera à étudier ensuite les différents types de flux, ainsi qu'à calculer le flux dans un corpus de manière automatique. Ensuite, nous présenterons dans la section 4.3 les différentes métriques basées sur le flux. En présentant chaque métrique que nous définirons, nous discuterons les hypothèses cognitives qui la sous-tendent. Enfin, pour savoir si la métrique que nous proposons prédit réellement la complexité syntaxique, nous évoquerons dans la section 4.4 l'évaluation des métriques du flux.

4.2 Représentation formelle

Rappelons tout abord que le flux de dépendance en une position donnée (entre deux mots d'une phrase) est l'ensemble des dépendances qui relient un mot à gauche de cette position à un mot à droite (Kahane, 2001). Nous représentons la position du flux de la façon suivante : (mot à gauche, mot à droite). Par exemple, dans la figure 43, la position du flux marquée par une ligne pointillée est représentée comme (*climate*, *risks*), le contenu de cette position du flux peut être représenté sous la forme d'une liste de relations {mitigate –*ccomp*-> risks, mitigate –*advcl*-> alleviating, climate <-*nmod*- risks}. Cependant, cette représentation a de nombreuses limites. Elle permet seulement de connaître la taille du flux, qui est le nombre de dépendances dans une position du flux. Pour pouvoir étudier la structure du flux et définir des métriques plus fines, nous avons besoin d'information sur les positions des sommets des différentes dépendances. La matrice de flux (Botalla, 2015) répond à ce besoin et fournit des informations plus riches. Dans

la section 4.2.1, nous allons d'abord présenter la matrice de flux. Cette dernière est conçue pour faciliter le stockage des informations du flux dans une base de données et permet le calcul de diverses métriques. Ensuite, dans la section 4.2.2 nous expliquerons des configurations de flux spécifiques, comprenant les dépendances en bouquet et les dépendances disjointes. Enfin, nous étudierons la difficulté de traitement de ces différentes configurations à la lumière des hypothèses cognitives.

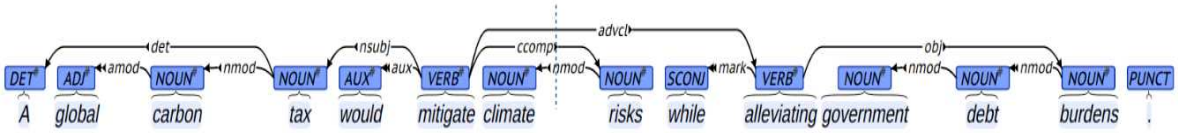

Figure 43 Position du flux (*climate*, *risks*) (UD_English-ParTUT, en_partut-ud-1239)

4.2.1 Matrice de flux

La matrice de flux est décrite ainsi dans Botalla (2014) :

> *« La matrice se présente sous la forme d'un tableau, dans lequel les lignes représentent la suite des mots (qui sont l'extrémité d'un élément du flux) à gauche de la position et les colonnes la suite des mots (qui sont l'extrémité d'un élément du flux) à droite. »*

La figure 44 montre le flux à la position (*the*, *victim*). Pour établir la matrice du flux, on numérote d'abord les mots à gauche du flux et les mots à droite en partant du mot le plus éloigné. À gauche de la position du flux, la numérotation se fait de gauche à droite. À droite, l'ordre est inverse, c'est-à-dire que la numérotation se fait de droite à gauche. Ainsi, la relation {the <-*det*-victim} peut être représentée dans le flux selon les positions du mot sous la forme en (numéro gauche, numéro droit) : (3, 2).

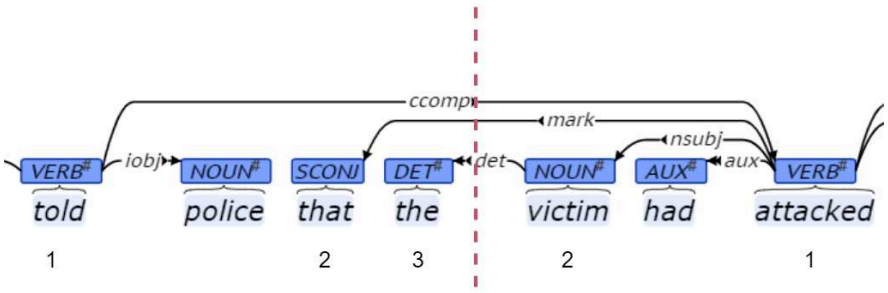

Figure 44 Représentation matricielle pour la position du flux (*the*, *victime*)

La figure 45 montre la configuration du flux (*the, victim*) et la matrice du flux dans le tableau 6. Dans la matrice du tableau 6, les lignes représentent les positions du mot à gauche et les colonnes représentent les positions du mot à droite. Ainsi, chaque élément non vide de la matrice contient une dépendance reliant un mot à gauche et un mot à droite. Par exemple, *told* est le premier mot à gauche, qui relie le premier mot *attacked* à droite : la relation {told – *ccomp*- > attacked} est ainsi à (1,1) dans la matrice.

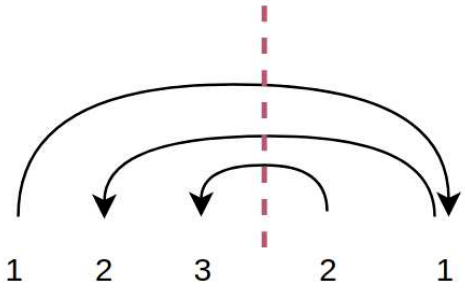

Figure 45 Configuration du flux pour la position (*the*, *victim*)

	1	**2**
1	(*told*, *attacked*) *ccomp*	
2	(*attacked*, *that*) *mark*	
3		(*victim*, *the*) *det*

Tableau 6 Matrice du flux pour la position (*the*, *victim*)

Cette formalisation permet de repérer les différentes configurations de flux dans le corpus.

Dans la section qui suit, on abordera les configurations du flux et leurs représentations matricielles.

4.2.2 Configurations possibles du flux

Les dépendances en bouquet

Les dépendances dans un flux qui partagent un même sommet forment une configuration en bouquet. Le sommet partagé du bouquet qui se trouve à droite de la position est un bouquet à gauche. Quant au cas du sommet partagé à gauche, il s'agit d'un bouquet à droite.

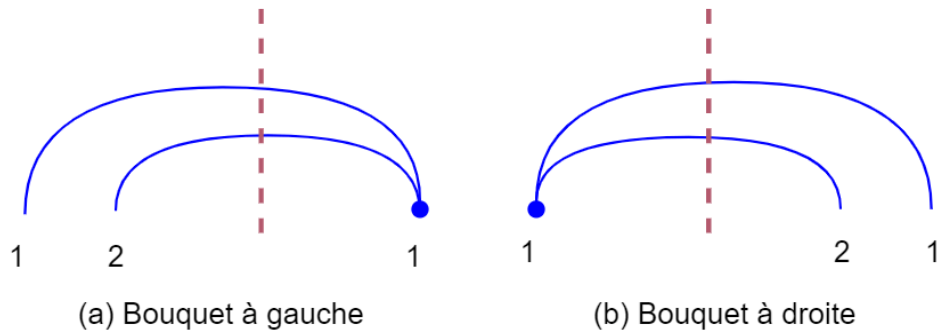

Figure 46 Bouquet à gauche et bouquet à droite

La figure 47 montre un sous-flux en bouquet et le tableau 7 sa représentation en matrice en couleur bleue. Nous pouvons observer que le bouquet à gauche se présente sous la forme d'une colonne, parce que le sommet partagé est à droite de la position du flux. Dans le cas du bouquet à droite, sa configuration dans la matrice prend la forme d'une ligne, puisque le sommet partagé se trouve à gauche de la position du flux.

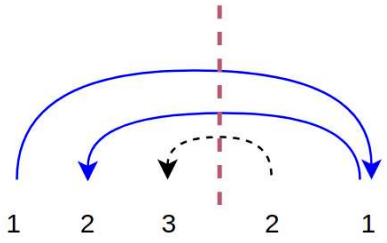

Figure 47 Configuration en bouquet dans la position (*the*, *victim*)

	1	**2**
1	(told, attacked), ccomp	
2	(attacked, that), mark	
3		(victim, the), det

Tableau 7 Représentation de bouquet dans la matrice pour la position (*the*, *victim*)

Les dépendances disjointes

Nous avons parlé plus haut du cas des dépendances dans un flux qui partagent un sommet. Mais il existe aussi des couples de dépendances dans un flux qui ne partagent aucun sommet commun. Nous dirons qu'elles sont ***disjointes***.

Dans la figure 48, la dépendance (3, 2) ne partage aucun sommet avec les deux autres dépendances (2, 1) et (1, 1). Ainsi, les dépendances (3, 2) et (2, 1) forment un sous-flux disjoint, tout comme les dépendances (3, 2) et (1, 1).

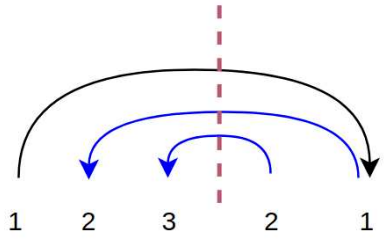

Figure 48 Dépendances disjointes dans la position du flux (*the*, *victim*)

Le tableau 8 montre la matrice du flux de la figure 48 et distingue les dépendances disjointes par différentes couleurs. Visiblement, les dépendances disjointes ne se trouvent dans aucune ligne ni colonne commune. Cela nous permet de déterminer facilement les dépendances disjointes dans la matrice.

	1	2
1	*(told, attacked), ccomp*	
2	*(attacked, that), mark*	
3		*(victim, the), det*

Tableau 8 Représentation matricielle des dépendances disjointes pour la position (*the*, *victim*)

Interprétation cognitive des configurations

Le fait d'avoir un sommet partagé dans le bouquet signifie que le nœud du sommet est étroitement lié avec les autres nœuds. Nous allons montrer comment la ***mise en facteur*** dans le cas de la présence de bouquet va faciliter le traitement et alléger la complexité syntaxique.

Nous considérons ici deux situations de mise en facteur respectivement pour le bouquet à droite et le bouquet à gauche. La première situation concerne le bouquet vers la droite, le nœud partagé arrive en premier. Comme ce qui est présenté dans la figure 49 : *a* et *b* sont liés par une dépendance, ainsi *a* et *b* forment-ils un syntagme *ab*. Ensuite, puisque *c* et *a* ont une dépendance, ils forment un syntagme *ac* ; *ac* et *ab* ont le facteur commun *a,* ils peuvent être factorisés en syntagme[39] : *a*(*b*·*c*). Et pour *d,* le processus est le même, l'on a finalement le syntagme factorisé *a*(*b*·*c*·*d*).

[39] Pour le syntagme *a*(*b*·*c*)*, b et c* ne forment pas en syntagme, ainsi, nous ajoutons un point entre *b* et *c* pour faire la distinction avec *bc* qui est un syntagme.

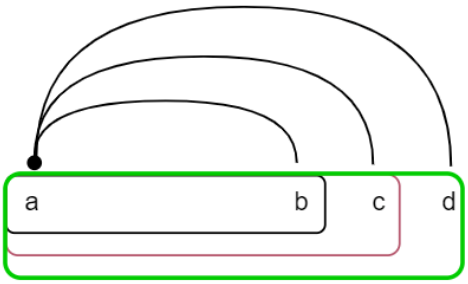

Figure 49 Mise en facteur pour un bouquet vers la droite

Il y a bien sûr un autre cas de mise en facteur, où le nœud partagé arrive en dernier. Par exemple, dans le cas de *le joli …*, *le* et *joli* n'ont pas de dépendance entre eux, mais *le* et *joli* seront liés avec un troisième nœud qui sera un nom. Ainsi, même si le troisième nœud n'est pas encore arrivé, puisque l'interlocuteur fait des prédictions pendant le traitement des phrases, ces deux nœuds peuvent former une amorce pour attendre leur nœud commun. Ceci est le cas dans le traitement de la structure en bouquet à gauche. Dans la figure 50, *a* et *b* attendent un nœud commun *d,* ainsi *a* et *b* peuvent-ils être factorisés en une amorce (*a*·*b*)_[40]. Ensuite, *c* attend le même nœud *d,* ainsi *c* et le groupe (*a*·*b*) peuvent-ils être factorisés en un groupe plus grand, (*a*·*b*·*c*). Et finalement, l'arrivée de *d* qui est le facteur commun permet de former le syntagme, (*a*·*b*·*c*)*d*.

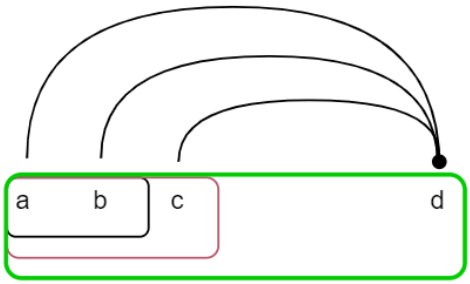

Figure 50 Mise en facteur pour un bouquet vers la gauche

Cependant, la situation de la construction ayant des dépendances disjointes est différent par rapport à celle en bouquet. La figure 51 montre un exemple ayant deux dépendances disjointes

[40] Comme le nœud commun *d* n'est pas encore arrivé, on le représente par _. C'est la place réservée au facteur commun des éléments entre parenthèses.

pour la position du flux (*très, belle*) : très <– *advmod* – belle, et une <-*det*- idée. *Une* et *très* ne peuvent pas être factorisés étant donné qu'ils ont une dépendance avec deux nœuds différents à gauche de leur position. Dans ce cas, nous ne pouvons pas faire de mise en facteur au moment de (*très, belle*), puisque aucun nœud n'est partagé et donc il n'y a pas de facteur commun.

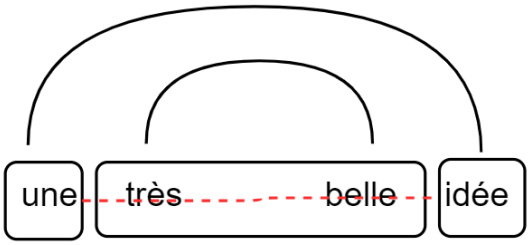

Figure 51 Exemple de deux dépendances disjointes

Ainsi, *une* et *très*, étant à gauche de la position du flux (*très*, *belle*), sont stockés dans la mémoire à court terme comme deux éléments indépendants. Donc, la configuration disjointe conduit à plus d'éléments à gérer dans la mémoire de travail.

De ce point de vue, la présence de dépendances en bouquet permet de compresser l'information dans la mémoire de travail par le processus de mise en facteur. Ainsi, la configuration en bouquet n'entraîne pas une augmentation aussi importante de la charge dans la mémoire de travail que des dépendances disjointes, puisque les éléments à gauche de la position du flux sont stockés indépendamment sans factorisation.

Les configurations de flux offrent de nombreux sujets d'étude intéressants pour la complexité syntaxique. Dans les sections suivantes, nous présenterons diverses métriques fondées sur le flux. En particulier, la section 4.3.2 discutera des métriques qui prennent spécifiquement en compte les configurations du flux.

4.3 Les métriques du flux et la complexité syntaxique

Cette section met l'accent sur les métriques du flux. Ces métriques du flux que nous allons

présenter découlent de différentes hypothèses sur la complexité syntaxique.

Les flux peuvent se concevoir de deux façons. D'une part, il s'agit du flux dans l'arbre de dépendance ordonné complet. Les métriques du flux élaborées à partir des arbres de dépendance complets sont des métriques du flux observé. Ceci sera évoqué tout d'abord dans les sections 4.3.1-4.3.3. D'autre part, nous pouvons déduire le flux en procédant mot par mot, auquel cas l'arbre de dépendance complet ne sera pas donné au début. Les informations dont nous disposons sont des mots et des dépendances déjà traitées. Le flux potentiel projectif et le flux requis sont deux métriques du flux élaborées et basées sur des arbres de dépendance partiels. Nous allons les présenter dans la section 4.3.4.

Nous remarquerons aussi que les possibilités utilisant le flux pour mesurer la complexité syntaxique ne sont pas toutes réalisables facilement. Conséquemment, nous choisirons les métriques du flux en fonction des hypothèses et de la faisabilité, et ce sont celles-ci que nous utiliserons dans les expériences (voir chapitre 5).

4.3.1 Taille du flux et minimisation de longueur de dépendance

Nous reprenons tout d'abord la taille du flux (voir aussi 0.2.3 dans l'introduction générale de cette thèse) proposée principalement dans les travaux de Jardonet (2009), Kahane, Yan et Botalla (2017) et Yan (2017). Notre apport va consister maintenant à mettre cette métrique en relation avec la longueur de dépendance. La figure 52 montre, en haut, la taille du flux pour toutes les positions et, en bas, la longueur de dépendance pour chaque mot. Nous pouvons voir que la longueur de la relation {tax -*det*-> A} est de 3, cela signifie également que cette relation est présente dans trois positions de flux : (*A, global*), (*global, carbon*), et (*carbon, tax*).

Si nous faisons la somme des tailles de flux et la somme des longueurs de dépendance pour cette phrase, on peut constater que ces deux valeurs sont identiques et égales à 21 (Kahane & Yan, 2019). Lorsqu'on fait la moyenne, les deux valeurs sont toutes deux à 1,75.

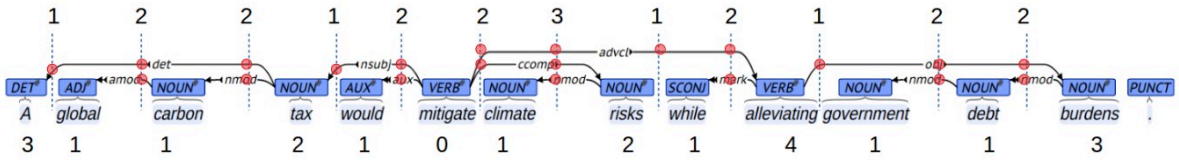

Figure 52 Exemple avec les tailles du flux et les longueurs de dépendance (UD_English-ParTUT, en_partut-ud-1239) (Kahane & Yan, 2019)

Longueur de dépendance de la phrase : 3 + 1 + 1 + 2 + 1 + 1 + 2 + 1 + 4 + 1 + 1 + 3 = 21
Taille du flux de la phrase : 1 + 2 + 2 + 1 + 2 + 2 + 3 + 1 + 2 + 1 + 2 + 2 = 21
Longueur de dépendance en moyenne = 21 / 12 = 1,75
Taille du flux en moyenne = 21 / 12 = 1,75

Grâce au calcul effectué dans l'exemple ci-dessus, nous constatons que le flux et la longueur de dépendance pour la phrase sont équivalents (Kahane & Yan, 2019). En effet, il s'agit d'une même métrique calculée de deux façons différentes.

Si nous utilisons le flux pour calculer la longueur de dépendance, nous calculons le nombre de positions du flux dont cette dépendance fait partie. Une dépendance qui se présente dans *n* positions du flux signifie qu'elle est de longueur *n*. Nous savons que la taille du flux d'une position inter-mots est le nombre de dépendances qui sont présentes dans cette position. Comme chaque dépendance dans une position du flux est présente une fois, la somme de toutes les tailles de flux dans la phrase est la somme du nombre de positions dont chaque dépendance fait partie. Cela nous mène à une conclusion importante : nous pouvons utiliser le flux pour réaliser le calcul de longueur de dépendance d'une phrase.

Interprétations cognitives

Au chapitre 3 dans la section 3.3.3.2, nous avons présenté la longueur de dépendance et la théorie de la minimisation de la dépendance (Liu, 2008 ; Futrell et al., 2015 ; Ferrer-i-Cancho, 2006 ; Gildea & Temperley, 2010). Comme la taille du flux totale est identique à la longueur de dépendance totale, la minimisation de la longueur de dépendance peut être interprétée en termes de flux.

Comme nous l'avons vu en introduction, le flux permet de saisir l'état de traitement au fur et à

mesure que chaque mot de la phrase est traité. L'ensemble de dépendances dans le flux est l'information nécessaire pour traiter avec succès le reste de la phrase. Comme il existe une limitation causée par la mémoire de travail pour traiter simultanément des informations (Cowan, 2001) (voir aussi chapitre 1, section 1.2), si le flux observé est très important dans une position inter-mots, il sera difficile à traiter. Ainsi, la minimisation de la longueur de dépendance peut être interprétée comme la tendance à avoir moins de dépendances dans chaque flux afin de ne pas traiter trop d'informations simultanément.

Nous savons qu'il existe également une limitation de la durée pour la mémoire de travail (Baddeley, 1992) (voir également le chapitre 1, section 1.2). Contrairement à l'interprétation du flux, la longueur de la dépendance est considérée comme une durée de temps pendant laquelle le mot doit rester dans la mémoire de travail, jusqu'à ce qu'il rencontre son dépendant ou son gouverneur (Hudson, 1995 ; Gibson, 2000). Ainsi, plus la dépendance est longue, plus elle doit rester longtemps dans la mémoire de travail. Pour une dépendance très longue, le mot serait détérioré dans la mémoire de travail et donc difficile à récupérer.

Pour conclure, les deux interprétations de la minimisation de longueur de dépendance s'intéressent à deux aspects différents de la limitation de la mémoire de travail. L'un est lié au traitement des informations concomitantes, l'autre est lié à celui du temps de stockage. Il y a d'autres points de vue possibles en termes de flux. La taille du flux ne considère que le nombre de dépendances, or il est également important de proposer des métriques qui prennent en compte différentes configurations du flux comme la configuration en bouquet et la configuration disjointe. Nous avons mentionné dans la section 4.2.2 la différence entre dépendances en bouquet et dépendances disjointes. Dans la section qui suit, nous aborderons des pistes pour étudier la complexité syntaxique en considérant ces diverses configurations du flux.

4.3.2 Étudier les configurations du flux

4.3.2.1 Configuration disjointe et auto-enchâssement

Nous avons introduit les configurations de flux dans la section 4.2.2 en utilisant les représentations syntaxiques abstraites, ce qui signifie qu'il n'y a pas d'informations spécifiques sur chaque nœud ni dépendance. La figure 53 montre une phrase artificielle analysée en dépendances (Yan, 2017). Nous pouvons remarquer que cette phrase contient une construction auto-enchâssée de niveau 3. Dans cette construction auto-enchâssée, il s'est produit des coupures pour former les deux syntagmes : *woman loves* et *man has*. De plus, chaque coupure tombe sur une dépendance entre verbe et sujet. En termes de flux, nous pouvons constater que, le flux pour la position (*child*, *knows*) contient trois dépendances *nsubj,* marquées en rouge, qui sont disjointes.

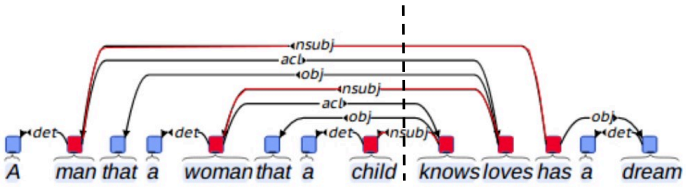

Figure 53 Auto-enchâssement à niveau 3 (Yan, 2017)

En effet, il a été montré que les dépendances disjointes correspondent à la construction auto-enchâssée (Kahane, Yan & Botalla, 2017 ; Yan, 2017). Comme nous l'avons mentionné dans les chapitres précédents, la limitation du niveau des structures auto-enchâssées est liées aux performances linguistiques, et de nombreux modèles de traitement des phrases dans le domaine de la psycholinguistique tentent d'expliquer cette difficulté. La capacité à interpréter les structures auto-enchâssées est essentielle si nous voulons utiliser le flux pour mesurer la complexité syntaxique. Cela renforce l'intérêt pour l'étude de la limitation des dépendances disjointes dans les flux.

Métrique : poids du flux

Kahane, Yan et Botalla (2017) et Yan (2017) ont proposé une métrique appelée *poids du flux*.

Définition : le poids du flux est le nombre maximal de dépendances disjointes, c'est-à-dire la taille maximale des sous-flux disjoints.

La figure 54 montre le poids pour chaque position du flux. Les flux en bouquet ont un poids de 1, car il n'y a pas deux dépendances disjointes. En ce qui concerne les positions (*the, victim*) et (*the, suspect*), le poids du flux est de 2 car il y a deux dépendances disjointes mais pas trois.

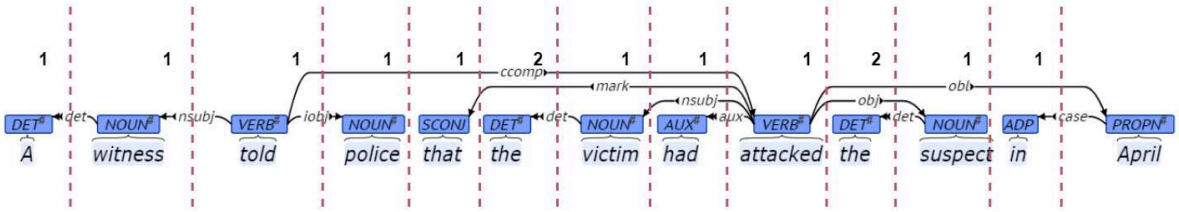

Figure 54 Exemple avec les poids du flux (UD_English-PUD, n01006011)

Comparé avec la taille du flux, le poids du flux prend en compte des informations sur la structure syntaxique car il tient compte de la relation configurationnelle entre les dépendances. Lorsque nous parlons de l'hypothèse cognitive du traitement des dépendances disjointes, nous savons que les éléments à gauche de la position du flux sont gardés dans la mémoire de travail indépendamment et sans factorisation possible. Car deux dépendances disjointes impliquent que la construction d'un syntagme est temporairement différée pour la construction d'un syntagme enchâssé. Ainsi, pour un flux de même taille, plus le poids du flux est important, plus la structure devrait être difficile à traiter à priori.

Dans la section précédente nous avons mentionné que la taille du flux se tend à être minimisée. Ceci nous a permis de proposer une explication à la théorie de la minimisation de longueur de dépendance. Une hypothèse similaire est faite sur le poids du flux. Cette hypothèse est que le poids est borné, puisque les niveaux auto-enchâssés dans la structure syntaxique sont limités (Lewis, 1996) (section 1.4.3.1). En d'autres termes, nous pensons qu'il y a également une tendance à réduire le poids pour chaque position du flux pour rendre le traitement des phrases plus économique.

Néanmoins, le poids ne peut remplacer totalement la taille dans le calcul de la complexité. Des positions du flux ayant un même poids peuvent en effet avoir des tailles très différentes. Par

exemple, dans la figure 54, beaucoup de positions ont un poids de 1, y compris lorsqu'il s'agit de bouquets dont les tailles sont différentes, comme pour les positions (*told*, *police*) ou (*had*, *attacked*). Selon les interprétations des configurations de flux discutées ci-dessus (section 4.2.2), les bouquets peuvent donner naissance à une mise en facteur, mais nous avons implicitement supposé que le nombre de nœuds n'a pas d'importance, ce qui n'est pas la réalité. Ainsi, une meilleure métrique devrait disposer d'informations sur la taille mais aussi sur le poids. Cela implique la nécessité de combiner ces deux notions dans une nouvelle métrique de complexité syntaxique. Nous allons nous focaliser sur cette problématique dans la partie suivante.

4.3.2.2 Combiner taille du flux et poids du flux

La taille et le poids du flux ne peuvent pas séparément expliquer entièrement la complexité syntaxique réelle. Nous pouvons faire mieux en combinant ces deux métriques dans un formule. Une des façons la plus simple est la combinaison linéaire[41].

La formule F6 montre l'expression pour une complexité syntaxique *C,* qui est une combinaison linéaire de la taille *S* et du poids *W*.

F6. $C = aS + (1 - a)W$

La figure 55 montre trois exemple (a), (b) et (c), nous pouvons voir que notre formule permet d'avoir des résultats pour *C* différenciés les uns des autres :

- Nous pouvons remarquer que (a) et (b) ont la même taille qui est de 3. Dans le cas du (a) de la figure 55, la complexité syntaxique C est : $3a + (1 - a)2 = a + 2$. Dans le cas du (b) de la figure 55, aucune dépendance ne forme un bouquet, ainsi les valeurs de la taille et du poids sont identiques. La complexité syntaxique C est : $3a +$

[41] Une combinaison linéaire de deux éléments est la somme des deux termes obtenus en multipliant chacun des éléments par une constante réelle. Par exemple, une combinaison linéaire de *x* et *y* serait une expression de la forme *ax* + *by*, où *a* et *b* sont des réels.

$(1 - a)3 = 3$.

- Deux positions du flux (a) et (c) ont le même poids qui est de 2. Nous pouvons également constater que C permet de différencier ces deux cas : pour (a), C est toujours $a + 2$; quand au (b), C est : $a(2 - 2) + 2 = 2$.

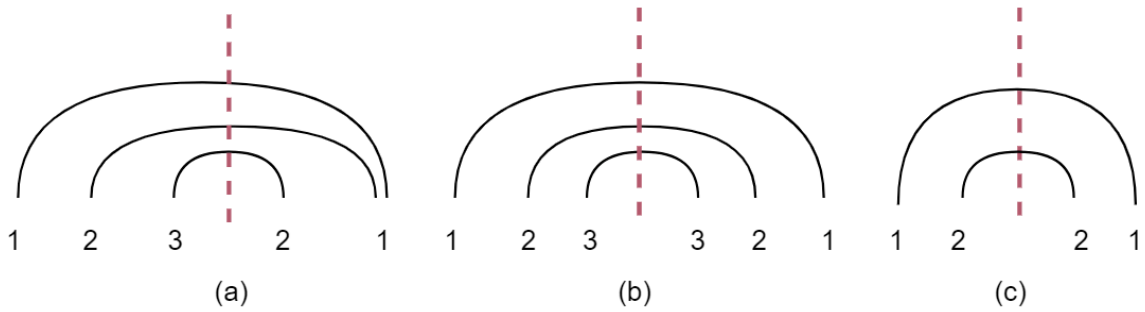

Figure 55 Trois configurations

La valeur de a dépend des hypothèses cognitives. Lorsque $a = 1$, la complexité C est identique à la valeur de la taille du flux. Le cas opposé est $a = 0$, où la complexité C est identique à la valeur du poids du flux. Bien entendu, nous pensons que la meilleure valeur de a se trouve entre 0 et 1. Pour trouver cette valeur, nous avons besoin de phrases dont la complexité syntaxique des positions inter-mots est déjà évaluée à l'aide d'expériences psycholinguistiques. Dans cette thèse, nous ne pouvons pas y arriver parce que nous n'avons pas disposé de données adéquates. Mais dans la section 4.4.3, nous allons expliquer comment obtenir des données attestées, afin de fournir une piste dans une future recherche sur le flux et la complexité syntaxique.

4.3.2.3 Étude des dépendances disjointes

Cette partie se concentre sur l'étude des dépendances disjointes. Nous avons précédemment constaté que les dépendances disjointes impliquent des constructions auto-enchâssées et que le poids du flux permet de mesurer le niveau d'auto-enchâssement.

En fait, il existe plus d'une hypothèse sur la difficulté de gérer une construction auto-enchâssée.

En parlant du poids du flux dans la section 4.3.2, nous avons montré que la difficulté est due à l'augmentation du stockage dans la mémoire de travail lors du traitement des dépendances disjointes. Nous prenons de nouveau l'exemple artificiel de la figure 56, où le flux marqué comprend trois dépendances de sujet disjointes. Comme le traitement des deux premières dépendances disjointes {man <-*nsubj*- has} et {woman <-*nsubj*- loves} est interrompu temporairement (voir aussi section 4.2.2), les deux sujets *man* et *woman* doivent être stockés dans la mémoire à court terme.

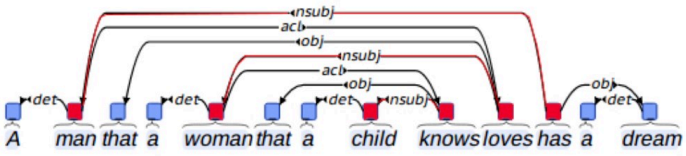

Figure 56 Auto-enchâssement à niveau 3 (Yan, 2017)

Cependant, il y a une autre hypothèse, liée à l'étude psycholinguistique sur la complexité syntaxique du chapitre 1, section 1.4.3.1, qui veut que la difficulté ne soit pas due à l'augmentation du stockage dans la mémoire de travail, mais à l'interférence causée par la similitude lorsqu'un nœud en traitement a plus d'un choix à gauche pour établir une seule dépendance. Précisément, au moment du traitement de *knows,* les trois sujets *child, man* et *woman* dans la phrase sont tous dans l'attente d'un verbe afin d'établir une dépendance de sujet. Le verbe *knows* est confronté à l'interférence causée par la similitude de ses candidats. Cela provoque un ralentissement du traitement, car il faut choisir le bon sujet parmi ces trois.

L'interférence a été mentionnée dans la section 1.4.3.2 du chapitre 1 pour expliquer la complexité de l'auto-enchâssement dans les travaux de Lewis et Vasishth (2005) et Lewis et al. (2006). Selon ces derniers, lorsque le premier verbe, *knows,* recherche son sujet, bien que des interférences augmentent la difficulté de recherche du sujet, le sujet ayant le plus haut degré d'activation dans la mémoire de travail sera privilégié. Comme la durée de stockage dans la mémoire à court terme est un facteur qui influence le niveau d'activation, le mot *child,* qui vient d'être traité, reste le plus activé par rapport à *woman* et à *man*. De ce fait, le sujet qui se propose est *child*. La même situation se produit dans le cas de *loves*. Comme *child* a déjà eu un verbe, il ne remplit donc plus les conditions pour être choisi : il nous reste *woman* et *man*. À

ce moment, le degré d'activation de *woman* est relativement élevé et la dépendance entre *woman* et *loves* sera établie.

Nous pensons que l'hypothèse ci-dessus peut être étudiée à travers le flux en examinant les dépendances disjointes. Nous pouvons prendre la dépendance de sujet et d'objet comme exemple.

La figure 57 montre quatre conditions des deux dépendances disjointes qui contiennent au moins une relation sujet (*nsubj* dans les treebanks UD) ou une relation objet (*obj* dans les treebanks UD).

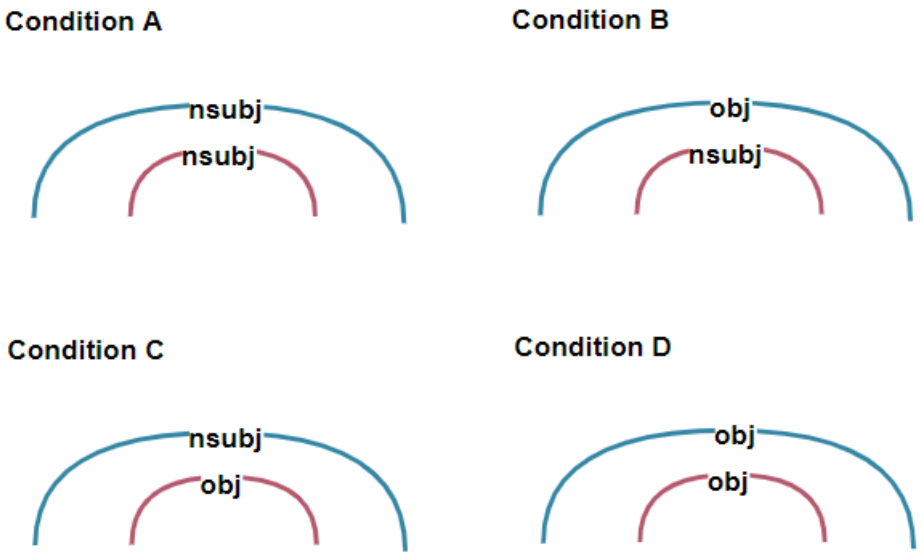

Figure 57 Quatres conditions pour deux dépendances disjointes (nous considérons ici seulement la dépendance du type objet (*obj*) et la dépendance du type sujet (*nsubj*))

Bas / ***Haut***	***nsubj***	***obj***
nsubj	$S_{bas}S_{haut}$	$S_{bas}O_{haut}$
obj	$O_{bas}S_{haut}$	$O_{bas}O_{haut}$

Tableau 9 Représentation des quatre conditions (*S: nsubj*, *O: obj*)

Comme la montre le tableau 9, nous pouvons représenter ces quatre conditions dans une

matrice, où les colonnes sont les dépendances inférieures dans la figure 57 et les lignes les dépendances supérieures.

Nous pouvons examiner les occurrences de ces quatre conditions dans les treebanks. Pour évaluer la distribution de ces quatre conditions, il nous faut les comparer avec les occurrences attendues.

Pour avoir les occurrences attendues pour ces quatre conditions, l'une des façons les plus simples d'y parvenir est d'utiliser la méthode probabiliste.

Tout d'abord, nous avons besoin du rapport *a* entre la probabilité de *nsubj* et la probabilité d'*obj*. Comme ce qui est exprimé dans la formule F7.

F7. $$a = \frac{P(nsubj)}{P(obj)}$$

Si les présentations de dépendance *nsubj* dans le flux disjoint sont des événements indépendants, nous pouvons obtenir la probabilité que se produisent les deux dépendances *nsubj* dans le flux disjoint par la formule F8 :

F8. $$\bar{P}(nsubj|nsubj) = P(nsubj) \times P(nsubj)$$

Avec la formule F7, on sait :

F9. $$P(nsubj) = a \times P(obj)$$

Ainsi, les occurrences attendues peuvent être calculées par les tableaux tableau 10 et tableau 11. Le tableau 10 montre les probabilités attendues pour ces quatre conditions et tableau 11 les nombres d'occurrences attendues :

Bas \ ***Haut***	***nsubj***	***obj***
nsubj	$P(obj\|obj) \times a^2$	$P(obj\|obj) \times a$
obj	$P(obj\|obj) \times a$	$P(obj\|obj)$

Tableau 10 Probabilités attendues

Bas / ***Haut***	***nsubj***	***obj***
nsubj	$O_{bas}O_{haut} \times a^2$	$O_{bas}O_{haut} \times a$
obj	$O_{bas}O_{haut} \times a$	$O_{bas}O_{haut}$

Tableau 11 Nombres d'occurrences attendues

Si nous trouvons que $S_{bas}S_{haut}$ est proche de ce qu'nous attendons, cela impliquera que les apparitions de deux dépendances sujet disjointes ne sont pas significativement rares dans le corpus. Par contre, si nous trouvons que $S_{bas}S_{haut}$ est beaucoup moins que ce que nous attendons, cela implique que la configuration $S_{bas}S_{haut}$ a tendance à être évitée. Dans ce cas, nous pensons que $S_{bas}S_{haut}$ crée probablement une plus grande complexité syntaxique. Une telle observation pourrait soutenir l'hypothèse présentée chez Lewis et Vasishth (2005) et Lewis et al. (2006).

Pour conclure, le flux nous permet d'examiner la complexité syntaxique des constructions auto-enchâssées dans deux directions. D'une part, la complexité pourrait être plus élevée lorsqu'il s'agit d'un poids plus important dans une position de flux. D'autre part, en regardant les dépendances disjointes, il nous permet de vérifier si la présence des dépendances disjointes similaires dans une position augmenterait la complexité syntaxique. Nous reprendrons cela dans la section 5.5.1 du chapitre 5.

4.3.3 Poids combiné

Le flux pour les positions inter-mot en même poids et taille peut comprendre des dépendances de différentes longueurs et nous pensons que cela peut jouer sur la complexité. Par exemple, dans la figure 58, deux positions du flux, (*told*, *police*) et (*attacked*, *the*), possèdent tous deux une taille de 2 et un poids de 1. De plus, ils sont aussi de configuration équivalente. C'est-à-dire que chaque flux contient deux dépendances, qui forment une configuration en bouquet. Un autre exemple dans la figure 59 est (*the*, *victim*) et (*the*, *suspect*), qui sont deux flux similaires

ayant la même taille et le même poids. Pour observer en détail les longueurs de dépendance dans ces positions, on observe les matrices des flux avec longueurs de dépendance pour la figure 58 dans le tableau 12 et le tableau 13, et pour la figure 59 le tableau 14 et le tableau 15.

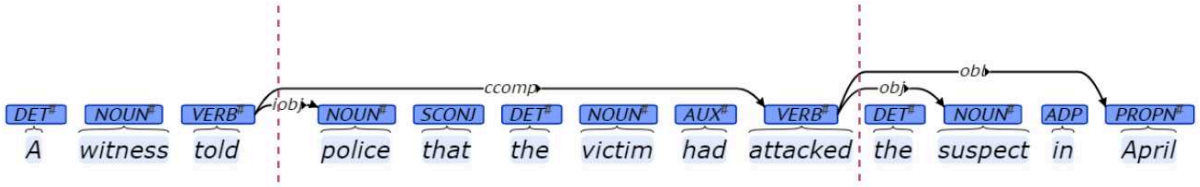

Figure 58 Position (*told*, *police*) et position (*attacked*, *the*) (UD_English-PUD, n01006011)

	1	2
1	{told- *ccomp*-> attacked} : **6**	{told -*iobj*-> police} : **1**

Tableau 12 Matrice du flux (*told, police*) avec les longueurs de dépendance.

	1	2
1	{attacked- *ccomp*-> April} : **4**	{attacked -*obj*-> suspect} : **2**

Tableau 13 Matrice du flux (*attacked, the*) avec les longueurs de dépendance.

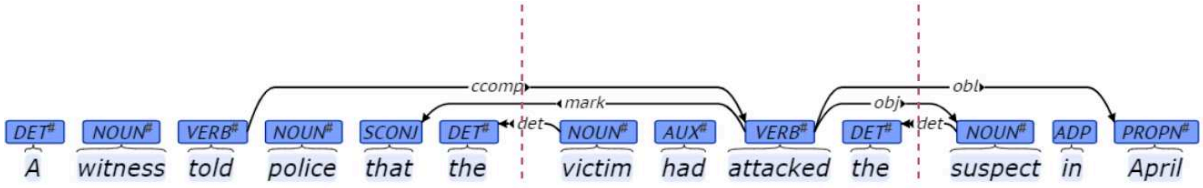

Figure 59 Position (*the*, *victim*) et position (*the*, *suspect*) (UD_English-PUD, n01006011)

	1	2
1	{told- *ccomp*-> attacked} : **6**	
2	{that <-*mark* - attacked} : **4**	
3		{the <- *det*- victim} : **1**

Tableau 14 Matrice du flux (*the, victim*) avec les longueurs de dépendance

	1	2
1	{attacked- *ccomp*-> April} : **4**	{attacked -*obj*-> suspect} : **2**
2		{the <-*det*- suspect} : **1**

Tableau 15 Matrice du flux (*the, suspect*) avec les longueurs de dépendance.

Nous pensons qu'il existe des différences au niveau de la complexité syntaxique dues à ses longueurs de dépendance différentes. Comme nous l'avons mentionné précédemment, la longueur de dépendance correspond à la durée pendant laquelle un nœud est stocké dans la mémoire de travail. En d'autres termes, plus longtemps un nœud est stocké dans la mémoire de travail, plus la détérioration de ce nœud est grave.

Yan et Kahane (2018) ont pris en compte les longueurs des dépendances disjointes pour calculer la complexité syntaxique. Dans leurs travaux, une possibilité proposée était de combiner la longueur de dépendance avec le poids du flux. La métrique de Yan et Kahane (2019) s'appelle le *poids combiné* (angl. *combined weight*).

Définition : le poids combiné dans une position donnée est la somme des longueurs des plus longues dépendances disjointes.

La figure 60 montre le poids combiné de chaque position du flux. Le flux pour la position (*the, victim*) et la position (*the, suspect*) contiennent chacun deux dépendances disjointes. Dans le flux pour la position (*the, victim*), les dépendances disjointes les plus longues sont : {told-*ccomp*->attacked} et {the<-*det*-victim}. Ainsi, le poids combiné est la somme des longueurs

de dépendance de ces deux dépendances, c'est-à-dire 6 + 1 = 7. S'il n'existe pas de seconde dépendance disjointe dans un flux, le poids combiné est justement la plus grande longueur de dépendance du flux.

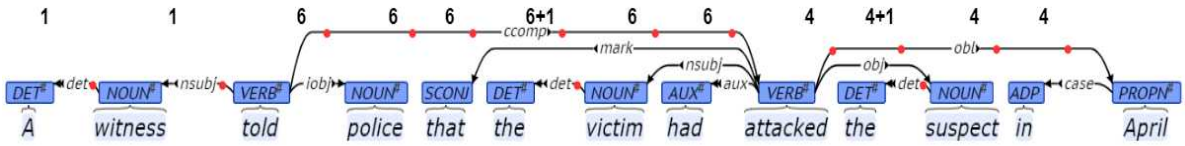

Figure 60 Valeurs du poids combiné (UD_English-PUD, n01006011)

En ce qui concerne le poids combiné total pour la phrase, il y a deux moyens de calculer sa valeur. La première façon est de faire la somme de toutes les valeurs du poids combiné pour chaque position du flux. La deuxième façon est de faire le calcul à partir des longueurs des dépendances disjointes pour chaque position du flux. Comme nous l'avons appris dans la section 4.3.1, la longueur d'une dépendance implique le nombre de positions inter-mots par lesquelles la dépendance passe. Pour l'exemple de la figure 60, les points rouges marquent les positions de flux par lesquelles passent les plus longues dépendances disjointes dans chaque position de flux. Par exemple, la dépendance *ccomp* passe par six positions inter-mots, et on peut constater que cette valeur 6 est calculée dans le poids combiné des six positions inter-mots où elle passe. Si nous additionnons tous les poids combinés, nous constatons que le poids combiné de la phrase est égal à la somme des carrés des plus longues dépendances disjointes :

Poids combiné pour la phrase :
1+1+6+6+6+(6+1)+6+6+4+(4+1)+4+4+4 = 56
= 1*1*(det)*+1*1*(nsubj)* +6*6*(ccomp)*+1*1*(det)*+4*4*(obl)*+1*1*(det)* = 56

La formule F10 exprime cette remarque. Le poids combiné représente en W_c . $L_{disjoint}$ représente dans toutes les positions du flux, les longueurs des plus grandes dépendances disjointes.

F10. $W_c \ pour\ la\ phrase = \sum {L_{disjoint}}^2$

Interprétations cognitives

Selon les sections 4.2.2 et 4.3.2, le poids du flux est lié au nombre de syntagmes dont le traitement est temporairement suspendu pour traiter des syntagmes enchâssés. Nous pensons que les longueurs maximales de dépendances disjointes sont liées au temps de suspension de chaque syntagme. Ainsi, l'hypothèse pour le poids combiné est qu'il représente non seulement une information sur la discontinuité de traitement du syntagme, mais aussi une information sur la durée de cette discontinuité. Dans la section 5.5.2 du chapitre 5, nous reviendrons sur les études quantitatives de cette métrique.

4.3.4 Flux potentiel (projectif) et flux requis

Les sections précédentes étaient des discussions sur des flux observés ; dans cette section, nous discutons des flux potentiels. L'analyse des flux observés se fonde sur l'arbre syntaxique ordonné complet, alors que l'analyse pour le flux potentiel projectif et pour le flux requis s'appuie sur un arbre syntaxique partiel, généralement la portion de structure qui précède la position inter-mot étudiée.

Métrique : Flux potentiel (projectif)

Kahane et Yan (2019) ont étudié le flux potentiel (voir la définition dans la suite) fondé sur le processus d'analyse de gauche à droite des phrases. Plus précisément, ce processus d'analyse fait référence à l'analyse automatique basée sur des transitions (version *Arc-Eager System*). Le calcul fait de plus l'hypothèse que le résultat de l'analyse est nécessairement un arbre projectif.

Pour mieux comprendre le flux potentiel projectif dans la suite, nous allons tout d'abord présenter l'analyse syntaxique basée sur des transitions. Selon Nivre (2003 ; 2008), un analyseur syntaxique basé sur des transitions consiste en deux structures de données, c'est-à-dire une pile et une file d'attente. La pile représente les *tokens* traités et encore accessibles, tandis que la file d'attente représente les *tokens* d'entrée restants (voir aussi la figure 61). Les actions de l'analyseur sont *Shift*, *Left_arc, Right_arc* et *Reduce,* que Nivre (2008) définit

comme suit.

Shift : pousser un mot dans la pile, c'est-à-dire déplacer un nœud de la file d'attente vers la pile. L'exemple de la figure 61 montre un processus *Shift*. Le mot *Il* est passé de la file d'attente à la pile.

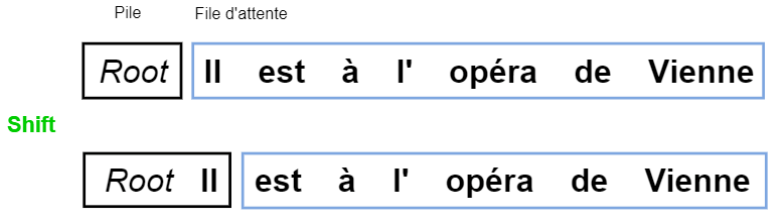

Figure 61 *Shift*

Left_arc : attacher le dernier nœud de la pile au premier nœud de la file d'attente, en créant un arc vers la gauche. Enlever le dernier nœud de la pile (puisque, étant gouverné, il n'est plus accessible si l'arbre est projectif). La figure 62 montre la création de l'arc vers la gauche entre *Il* et *est*. Le dernier nœud de la pile *Il* est retiré (voir aussi le tableau 16).

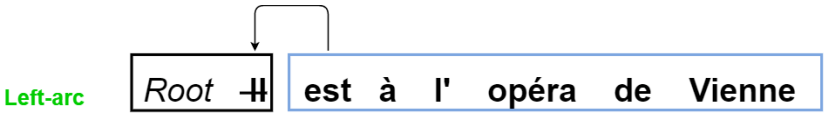

Figure 62 *Left-arc*

Right_arc : attacher le dernier nœud de la pile au premier nœud de la file d'attente, en créant un arc vers la droite. Pousser le premier nœud de la file d'attente dans la pile.

La figure 63 montre le processus de *Right-arc*. Les mots *est* et *à* se trouvent respectivement en dernier nœud de la pile et premier nœud de la file d'attente. L'analyseur attache *est* et *à* en créant un arc vers la droite, où *est* est le gouverneur et *à* le dépendant. Ensuite, l'analyseur déplace *à* de la file d'attente dans la pile (voir aussi le tableau 16).

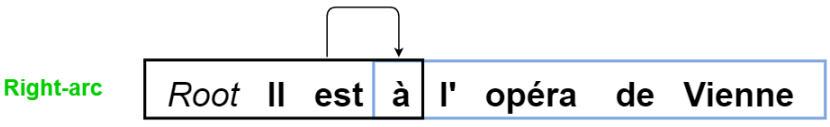

Figure 63 *Right-arc*

– *Reduce* : supprimer le dernier nœud de la pile à condition qu'une tête lui ait déjà été attribuée dans la structure des dépendances. Selon Nivre (2008) :

> *"This transition is needed to remove a node previously pushed onto the stack in a Right-Arc transition after it has found all its own right-dependents."*

La figure 64 montre quatre applications successives de *Reduce* au moment du traitement d'*aujourd'hui. Vienne*, *de*, *opéra*, et *à* qui sont progressivement retirés de la pile avant de passer dans une action de *Right-arc* entre *aujourd'hui* et *est,* parce que ces quatre tokens dans la pile ont trouvé leur tête et tous leurs dépendants à droite.

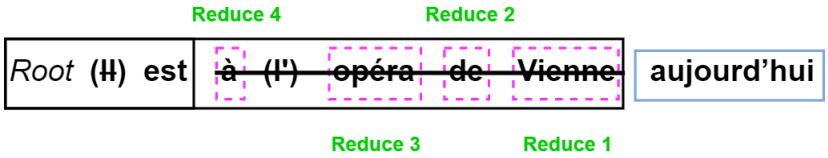

Figure 64 *Reduce*

Le traitement des phrases consiste à réduire la pile en gardant seulement la racine (*root*) et à vider la file d'attente. Le tableau 8 montre chaque étape de l'analyseur basée sur des transitions. Pour les actions de *Left-arc* et de *Right-arc,* l'analyseur examine le dernier nœud de la pile et le premier nœud de la file pour la création de dépendances.

Nous pouvons constater que, à chaque nouvelle dépendance établie, les nœuds restés dans la pile sont ceux qui ont potentiellement un dépendant dans la file d'attente. Dans le cas d'une création de la dépendance vers la gauche, le dernier nœud de la pile est le dépendant. En maintenant le principe de projectivité, ce nœud ne peut pas avoir un dépendant dans la file d'attente pour établir une nouvelle dépendance. Ainsi, le dépendant doit être retiré de la pile. En ce qui concerne la création de la dépendance vers la droite, le gouverneur et le dépendant peuvent tous être la tête d'un nœud dans la file d'attente. Par conséquent, les deux doivent être dans la pile.

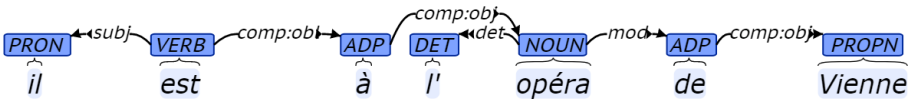

Figure 65 Exemple

Action	Pile	File d'attente	Dépendance ajoutée
-	[ROOT]	[il, est, à, l', opéra, de, Vienne]	
Shift	[ROOT, il]	[est, à, l', opéra, de, Vienne]	
Left-arc	[ROOT]	[est, à, l', opéra, de, Vienne]	{il <–*subj*- est}
Right-arc	[ROOT, est]	[à, l', opéra, de, Vienne]	{ROOT -*root*-> est}
Right-arc	[ROOT, est, à]	[l', opéra, de, Vienne]	{est -*comp:obl*-> à}
Shift	[ROOT, est, à, l']	[opéra, de, Vienne]	
Left-arc	[ROOT, est, à]	[opéra, de, Vienne]	{l' <-*det*- opéra}
Right-arc	[ROOT, est, à, opéra]	[de, Vienne]	{à -*comp:obj*-> opéra}
Right-arc	[ROOT, est, à, opéra, de]	[Vienne]	{opéra -*mod*-> de}
Right-arc	[ROOT, est, à, opéra, de, Vienne]	[]	{de -*comp:obj*-> Vienne}
Reduce	[ROOT, est, à, opéra, de]	[]	
Reduce	[ROOT, est, à, opéra]	[]	
Reduce	[ROOT, est, à]	[]	

Reduce	[ROOT, est]	[]	
Done	[ROOT]	[]	

Tableau 16 Actions de l'analyse

Du point de vue cognitif, les mots de la pile doivent être stockés dans la mémoire de travail, prêts à former de nouvelles dépendances avec les mots entrants. Les mots dans la file d'attente n'ont pas besoin d'être stockés car ils ne sont pas encore arrivés. Nous pouvons essayer d'observer la taille de la pile à chaque étape pour étudier la complexité syntaxique. Kahane et Yan (2019) ont proposé une métrique fondée sur le flux pour l'étudier, c'est le flux potentiel. Le flux potentiel enregistre l'ensemble des mots qui restent accessibles en mémoire lors de chaque traitement. La définition suivante du flux potentiel est donnée par Kahane et Yan (2019).

Définition :

"We call potential flux in a given inter-word position the set of words before the position which are likely to be linked to words after it."

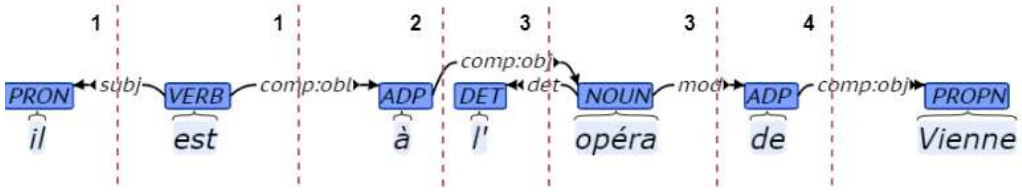

Figure 66 Valeurs du flux potentiel

Le flux potentiel de l'exemple précédent est illustré dans la figure 66. Par exemple, sur la position (*de*, *Vienne*), le flux potentiel est de 4, ce qui signifie qu'il y a quatre nœuds sur la gauche qui peuvent construire de nouvelles dépendances avec un nœud sur la droite : *est*, *à*, *opera* et *de*. Cet ensemble des nœuds correspond à la pile avant l'ajout de {*comp:obj*} dans l'analyse basée sur des transitions (voir le tableau 16).

Le flux potentiel considère que, lorsque les phrases sont traitées de gauche à droite, les informations à propos des mots qui ne sont pas encore présents sont inconnues. C'est la situation qui se présente dans le cas d'un traitement des phrases spontané. L'interlocuteur

n'attend pas que tous les mots soient arrivés pour commencer le traitement de la structure.

Nous pensons que les interlocuteurs font des prédictions sur les dépendances pendant le traitement. Les nœuds qui ont le potentiel d'établir une dépendance avec les mots suivants sont temporairement stockés dans la mémoire de travail. Les flux potentiels enregistrent l'ensemble des nœuds qui doivent être stockés lors du traitement d'une phrase. En raison de la limitation de la mémoire de travail, il peut également y avoir une certaine limitation au flux potentiel.

De plus, en respectant le principe de la projectivité, nous constatons que le flux potentiel n'est pas le même après la création d'une dépendance vers la gauche qu'après celle vers la droite. Lorsqu'il s'agit d'une création d'une dépendance vers la droite, le dépendant se trouve à droite de son gouverneur. Ils peuvent tous les deux avoir un dépendant à droite et la construction reste projective. Par contre en ce qui concerne la création d'une dépendance vers la gauche, seul le gouverneur peut avoir un dépendant à droite sans créer de non-projectivité (voir aussi la figure 67 pour la différence entre ces deux cas). C'est pourquoi les deux nœuds *Il* et *l'* ne peuvent pas faire partie du flux potentiel de la position (*de*, *vienne*), puisqu'ils sont dépendants à gauche et que construire pour eux de nouvelles dépendances avec un nœud à droite ne respecte pas la projectivité.

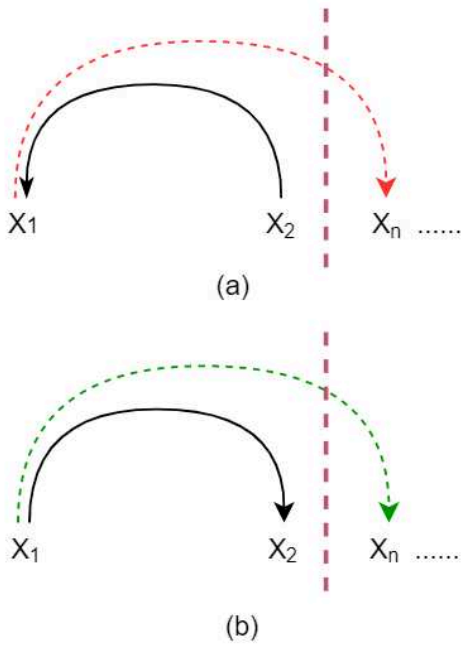

Figure 67 Dépendance vers la gauche et dépendance vers la droite (entre X_1 et X_2)

Enfin, nous pensons qu'il y a une asymétrie dans la distribution du flux potentiel au niveau de la direction de dépendance, et que les structures avec le plus de relations de dépendance vers la droite ont des tailles du flux potentiel plus importantes que celles avec le plus de relations de dépendance vers la gauche. Dans le chapitre 5 section 5.4, nous allons étudier le flux potentiel dans les treebanks SUD.

Métrique : Flux requis

Dans 3.3.3.3, nous avons présenté la densité de dépendance (angl : *dependency density*) de Hudson (1995). Rappelons que la densité de dépendance est le nombre de dépendances « ouvertes » à un moment donné. Comme la densité de dépendance ne considère que les dépendances ouvertes qui sont requises, nous utilisons le terme *flux requis* pour la notion de densité de dépendance de Hudson (1995).

Pour le flux potentiel, un nœud de la partie gauche en fait partie tant qu'il peut établir une dépendance avec un nœud de la partie droite en respectant le principe de la projectivité. Le flux requis, lui, prend en compte non seulement le principe de la projectivité, mais filtre également les dépendances potentielles établies entre les nœuds à gauche et ceux à droite, ne laissant que

les dépendances jugées nécessaires pour former une phrase.

Pour déterminer la valeur du flux requis systématiquement dans les *treebanks,* il faut compter le nombre de dépendances ouvertes pour chaque position inter-mots. Tout d'abord, il s'agit de toutes les dépendances vers la droite, car le dépendant est déjà traité et attend encore son gouverneur. Nous savons que, dans l'arbre de dépendance, chaque dépendant doit avoir son gouverneur ; si son gouverneur n'est pas encore arrivé à un moment donné, cette dépendance est incomplète.

Ensuite, il s'agit des dépendances vers la gauche qui sont nécessaires pour former une phrase. Cela concerne les dépendances vers les compléments postposés. Pour un mot dont le gouverneur a déjà été traité, seule la dépendance vers son complément obligatoire dans la suite est certaine. Dans ce cas, le mot attend son complément pour former une phrase, ce qui crée également une dépendance incomplète.

La figure 68 est un exemple extrait du *treebank SUD-French_Spoken*. Nous donnons les valeurs du flux requis et de la taille du flux pour chaque position inter-mots :

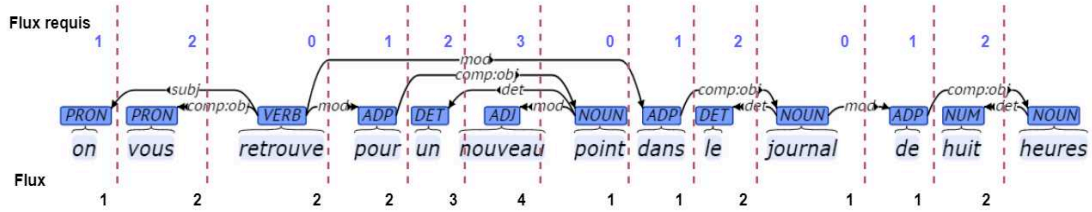

Figure 68 Valeurs du flux requis et valeurs du flux observé dans un arbre de dépendance (SUD_French_Spocken)

Par exemple, la taille du flux pour (*nouveau*, *point*) est de 4, car il contient quatre dépendances :

> {nouveau <-*mod*- point}, {un <-*det*- point}, {pour -*comp:obj*-> point} et {retrouve -*mod*-> dans}.

Le flux requis est de 3, car il s'agit de trois dépendances obligatoires qui sont incomplètes au moment du traitement du mot *nouveau* : {nouveau <-*mod*- point} et {un <-*det*- point} sont deux dépendances à gauche, *nouveau* et *un* attendent leur gouverneur *point* dans la suite. {pour

<-*comp:obj*- point} est aussi considéré comme une dépendance requise, car le complément de *pour* est obligatoire. Nous pouvons aussi constater que les dépendances à droite du type modifieur *mod* ne font pas partie du flux requis : la dépendance {retrouve -*mod*-> dans} n'est pas considérée comme une dépendance incomplète au moment du traitement de *nouveau*, car *dans le journal* est un modifieur et non un complément obligatoire.

La similarité entre un flux requis et un flux potentiel est qu'ils s'appuient tous deux sur l'arbre non complet. Cet arbre partiel est la partie gauche de cette position qui contient les nœuds déjà analysés. Les deux métriques sont de nature à prédire la partie droite de la position, qui n'est pas encore arrivée. Ainsi, nous pensons que comme le flux potentiel, les valeurs de tailles du flux requis pourraient également avoir des contraintes.

Par ailleurs, on pourrait observer dans les treebanks l'asymétrie de la distribution du flux requis au niveau de la direction de dépendance. Une structure avec plus de relations de dépendance vers la gauche a des tailles du flux requis plus importantes qu'une structure avec plus de relations de dépendance vers la droite. En effet toutes les dépendances vers la gauche sont des dépendances requises, (qui doivent être établies pendant le traitement de la phrase de gauche à droite). Mais pour les dépendances vers la droite, seule la dépendance dont le dépendant à droite est un complément postposé est considéré comme une dépendance requise. Dans la section 5.5.3, nous présenterons nos résultats et analyses du flux requis dans les treebanks SUD.

4.4 Méthodes d'évaluation des métriques du flux

Après les explications sur les métriques du flux, une question se pose : comment savoir si la métrique est effectivement liée à la complexité syntaxique ? Tout d'abord, on peut utiliser une approche quantitative pour faire les évaluations des métriques du flux. C'est la solution que nous avons utilisée dans les expériences des métriques du flux, qui seront présentés dans le chapitre 5. Nous expliquerons cette solution plus en détail dans la section 4.4.1.

De plus, la solution optimale serait qu'on puisse comparer la complexité syntaxique réelle

évaluée à des expériences psycholinguistiques avec les résultats obtenus par les métriques du flux. Ainsi, nous allons présenter les différentes façons pour évaluer la complexité syntaxique réelle, à savoir les évaluations par des méthodes intuitives (section 4.4.2), et les évaluations par des expériences psycholinguistiques (section 4.4.3). L'objectif de cette partie est d'évoquer des informations préliminaires afin d'élaborer des études sur le flux et la complexité syntaxiques après cette thèse.

4.4.1 Méthode quantitative

Cette section se concentre sur l'évaluation quantitative des métriques sans paramètre comme nous l'avions inséré dans la combinaison taille-poids. L'évaluation quantitative peut être divisée en deux approches. D'une part, on peut évaluer les métriques en analysant la distribution de ses valeurs obtenues dans les treebanks. Par exemple, nous pouvons examiner la rareté des valeurs suffisamment grandes, car les constructions très complexes ont tendance à être évitées dans les données réelles. D'autre part, il s'agit de proposer un modèle de base (angl. *baseline*) comme le contexte aléatoire, et comparer les résultats des métriques dans le modèle de base avec ceux obtenus dans les treebanks originaux.

Analyser les valeurs métriques en fonction de la rareté

Liu (2008) et Futrell et al. (2015) ont calculé la longueur de dépendance dans les treebanks de différentes langues. Selon leurs travaux, les dépendances très longues sont rares (voir aussi section 3.3.3.2 du chapitre 3). Dans cette thèse, nous allons suivre également cette approche pour évaluer les métriques du flux (sans paramètre) (voir chapitre 5). Notre travail sera développé avec l'hypothèse que la limitation de la mémoire de travail se manifesterait dans le traitement syntaxique des phrases. Dans le corpus, on ne trouverait pas fréquemment les constructions nécessitant une plus grande charge de mémoire de travail en raison de la plus grande difficulté de traitement, et il n'y aurait pas de constructions complexes dépassant la capacité de la mémoire de travail. Ainsi, un des travaux essentiels est d'observer la distribution des valeurs de complexité obtenues par les métriques, et d'examiner si les valeurs de complexité

obtenues par les métriques sont bornées.

Comparer les valeurs métriques avec celles du modèle de base (angl. *baseline*)

Dans cette approche, un ou plusieurs modèles de base peuvent être conçus en fonction de l'hypothèse à tester. Par exemple, pour tester que les valeurs d'une métrique de complexité syntaxique dans le langage naturel ont une distribution différente de celle dans les séquences en ordre aléatoire, on peut concevoir un treebank contenant des séquences ayant l'ordre des mots aléatoire comme modèle de base. Ensuite, les métriques sont calculées dans le treebank original et dans le modèle de base, et les résultats des deux modèles sont analysés.

Cette méthode est généralement utilisée dans les études sur la minimisation de longueur de dépendance (voir chapitre 3, section 3.3.3.2), comme celles de Liu (2008), Futrell et al. (2015), Gildea et Temperley (2010), et Ferrer-i-Cancho (2006). L'hypothèse est que les modèles de langue configurent l'ordre des mots afin de réduire la complexité syntaxique et de rendre le traitement des phrases plus simple. Selon leurs études, les longueurs de dépendances sont beaucoup plus grandes dans les séquences en ordre aléatoire. Cela signifie que l'interlocuteur préfère l'ordre de mots avec des longueurs de dépendance plus courtes et tend donc à minimiser la longueur de dépendance.

La première approche tend à trouver des limitations aux valeurs métriques et vérifie qu'elles existent dans les treebanks de différentes langues. L'approche ayant le modèle de base tend à voir si la métrique renvoie à des caractéristiques du langage naturel qui impliquent une complexité syntaxique plus importante. Nous pensons que ces deux approches sont complémentaires, en particulier lorsque la limitation des valeurs de la métrique n'est pas clairement déterminée.

Commentaires

Notre position est qu'il y a deux avantages à utiliser des méthodes quantitatives pour évaluer la métrique. Tout d'abord, les évaluations quantitatives sont effectuées à l'aide de données linguistiques réelles plutôt que de phrases artificiellement construites. Deuxièmement, les

évaluations quantitatives sont moins coûteuses à réaliser. Elles peuvent en effet être reproduites autant que nécessaire une fois que des données annotées de bonne qualité sont disponibles, sans l'intervention d'instruments ou de personnel supplémentaires.

Cependant, l'évaluation quantitative nécessite que la complexité syntaxique obtenue par la métrique puisse être calculée directement à partir des données, telles que la taille du flux, le poids, le poids combiné, le flux potentiel et le flux requis. Dans la section 4.3, nous avons également proposé la formule F6 pour calculer une complexité *C,* qui utilise un paramètre *a*. Dans ce cas-là, il peut être intéressant de faire des expériences psycholinguistiques pour ajuster au mieux les paramètres et obtenir les valeurs de la complexité réelle (voir section 4.4.3).

4.4.2 Méthode intuitive

Au chapitre 2, nous avons expliqué que selon Chomsky (1965), la performance linguistique peut être évaluée en termes d'acceptabilité. En d'autres termes, une phrase dont la structure syntaxique est très complexe sera moins acceptable, et certaines structures trop complexes seront inacceptables.

La méthode intuitive consiste à faire évaluer par le participant l'acceptabilité d'une phrase en fonction de sa propre intuition. Les premiers travaux, tels que ceux de Holmes et Forster (1973), ont proposé la notation de l'acceptabilité — aussi dite naturalité (angl : *naturalness*) — des phrases, comme l'une des nombreuses méthodes d'évaluation.

Par exemple, pour un ensemble de phrases qui contiennent une proposition relative en ajout périphérique droit [42], et de phrases auto-enchâssées [43], les participants devaient noter

[42] Voici un exemple de phrase relatives en ajout périphérique droite de Holmes et Forster (1972) :
Ann was wearing the pretty frock that her mother made.

[43] Un exemple de phrase avec auto-enchâssement de Holmes et Forster (1972) :
The truck that Bill was driving crashed into the post.

l'acceptabilité de la façon suivante[44] :

> *"A five-point scale was used, ranging from +2, 'perfectly natural and acceptable,' to -2, 'very unnatural (clumsy or bizarre)'"*

De cette manière, les phrases auto-enchâssés ont obtenu une acceptabilité moyenne de 1,25 contre 1,14 pour les phrases avec relatives en périphérie droite.

Holmes et Forster ont comparé la moyenne d'acceptabilité avec le nombre de mots correctement rapportés (voir en détail dans 4.4.3 sur la *méthode psychologique*) (angl : *number of words correctly reported*) pour les phrases avec différents types de structures. Selon Holmes (1973) :

> *"One important variable is the judged "naturalness" of the sentences. Holmes and Forster (1972) found that syntactic effects on comprehension difficulty could be completely obscured if naturalness was not held constant over conditions. In the present experiment, despite every attempt to make all the sentences natural, a significant correlation between naturalness and perceptual complexity was obtained."*

La méthode intuitive permet d'obtenir rapidement des jugements humains sur la complexité des phrases. Cependant, l'évaluation de l'acceptabilité est souvent influencée par des facteurs autres que la syntaxe, comme la sémantique. En outre, cette méthode évalue la perception globale de la phrase et ne nous aide pas à comprendre les détails de la difficulté de traitement de la phrase. Enfin, elle n'est pas assez objective et les réponses des participants ne sont pas toujours homogènes.

[44] Les phrases en question sont de même longueur pour éviter l'influence de la longueur de phrase sur la complexité syntaxique. En fait, de plus en plus d'études utilisent aujourd'hui de véritables phrases issues de corpus, nous ne rentrons pas en détail sur les conclusions concernant la complexité syntaxique avec ces exemples.

4.4.3 Méthode psychologique

La méthode psychologique se concentre sur l'analyse du comportement que les gens adoptent en réponse à des stimuli. Par rapport à la méthode intuitive, cette approche réduit l'influence de la subjectivité. Afin d'obtenir des valeurs de la complexité syntaxique et d'effectuer les évaluations de métrique, les méthodes varient en fonction de la manière dont les données expérimentales sont obtenues.

Nombre de mots correctement rapportés (angl. *number of words correctly reported*) :

Forster (1970) est l'un des premiers travaux à avoir proposé d'enregistrer le nombre de mots correctement rapportés pour évaluer la complexité syntaxique. La technique utilisée dans ses expériences s'appelle *Présentation Visuelle en Série Rapide* (angl. *Rapid Serial Visual Presentation*) qui est décrite par Holmes & Forster (1972) :

> *"Briefly, this involves visually presenting the words of the sentence in rapid succession. Each word is super-imposed on the preceding word, which prevents any cumulative sensory storage of the input. These forces S (subject) to process the input at the same rate as it is presented. One advantage of this technique is that usable error rates can be achieved with normal sentences, without having to resort to the highly abnormal doubly self-embedded constructions used by Fodor et al. (1968)."*

Selon Holmes et Forster, le nombre de mots correctement rapportés permet d'évaluer la complexité syntaxique pour le destinataire. Homes et Forster pensent également que l'endroit où il s'est produit des erreurs (par exemple : un mot manqué ou mal placé) implique éventuellement un traitement syntaxique plus difficile. Un autre avantage de cette méthode est que sa procédure est simple et qu'elle n'a pas besoin d'équipements spécifiques. Un de ses inconvénients est qu'il n'est pas possible d'exclure l'influence de la sémantique sur leurs résultats.

Méthode d'oculométrie

La méthode d'oculométrie consiste à analyser les données des mouvements oculaires (angl. *eye-tracking*). L'hypothèse est que la durée que le regard d'une personne pose sur chaque élément de la phrase pendant la lecture d'une phrase n'est pas réparti de manière égale. Plus la durée de fixation est longue, plus il est probable qu'il y ait eu des difficultés dans le traitement.

De nos jours, les études des mouvements oculaires sur le traitement syntaxique se divisent en deux catégories, l'une est le temps de lecture qui est obtenu en enregistrant la durée pendant laquelle le regard reste dans une zone (cette zone peut être un groupe de mots, ou un mot). L'autre catégorie est l'étude des saccades oculaires régressives (*eye back-tracking*) (Holmes et O'Regan, 1981 ; Staub, 2010), lorsque les gens éprouvent des difficultés de traitement et de l'incertitude en redirigeant leur regard vers une zone qu'ils ont déjà lue. Holmes et O'Regan (1981) ont d'abord tenté d'obtenir des données sur la complexité syntaxique à partir de données des mouvements oculaires. Dans leurs travaux, les relatives avec l'extraction du sujet et celles avec l'extraction de l'objet (voir aussi chapitre 1 section 1.4.1) en français ont été étudiées. Les deux auteurs pensent que ces deux types de mouvements oculaires concernent les problèmes du traitement des phrases dans des situations différentes, et que la régression permet de détecter la difficulté au niveau du traitement de la structure profonde :

> *"Subject relatives were read with the least amount of regressive behavior compared with both types of object relative[45]. Differences between the two kinds of object relative in the starting points of the regression sequences suggest also that it is the particular combination of deep and surface structure properties which is important in determining when the eye*

[45] Les exemples de Holmes et O'Regan (1981) :
Relative avec l'extraction du sujet : *Je crois que le sauvage qui va attaquer le chasseur monte sur un cheval noir.*
Relative avec l'extraction de l'objet (1) : *Je crois que le sauvage que le chasseur va attaquer monte sur un cheval noir.*
Relative avec l'extraction de l'objet (2) : *Je crois que le sauvage que va attaquer le chasseur monte sur un cheval noir.*

regresses"

"The difference between subject and object relatives in overall amount of relooking suggests that regressive movements occur more often in structures which are psychologically more complex."

Nous pensons que les données sur les mouvements oculaires peuvent être utilisées pour évaluer les résultats obtenus par les métriques. D'une part, il s'agit de la comparaison entre des données sur la durée du regard et des résultats de la complexité syntaxique calculée par la métrique du flux. D'autre part, on peut comparer les résultats avec ceux de données sur la probabilité de régression oculaire.

Par rapport aux méthodes présentées précédemment, la méthode de mouvements oculaires nécessite plus d'équipement expérimental et éventuellement plus de travail de traitement des données. Mais elle peut enregistrer la façon dont une personne traite naturellement le langage et donner des informations plus riches.

Méthode du *Self-paced reading* (SPR)

Les études sur le traitement des phrases dans le domaine de la psycholinguistique présentées au chapitre 1 utilisent souvent cette méthode (voir aussi King et Just (1991), Gibson (1998, 2000), Grodner et Gibson (2005), et Levy (2008)). La procédure principale de cette technique est de laisser le participant lire la phrase mot par mot (ou groupe de mot) à son propre rythme. Chaque fois, l'écran affiche uniquement un mot. Le participant doit appuyer sur un bouton pour le masquer et afficher le mot suivant. L'intervalle entre chaque appui du bouton est enregistré et il représente le temps de lecture du mot affiché à ce moment précis.

Comme la méthode des mouvement oculaires, cette méthode permet d'avoir le temps de lecture de chaque mot dans une phrase, et donc elle permet d'évaluer les résultats obtenus par les métriques pour chaque mot. C'est un avantage comparé avec la méthode utilisant le nombre de mots correctement rapportés, car avec cette dernière on peut obtenir l'information de la complexité seulement là où il s'est produit des erreurs. D'ailleurs, cette méthode est également

relativement facile à mettre en œuvre. Aujourd'hui, avec un outil informatique (par exemple, psychoPy3, bibliothèque de Python) on peut créer des expériences qui permettent aux participants de les faire via l'outil et de les évaluer donc automatiquement.

Comparant avec la méthode d'oculométrie, Clifton et Staub (2011) pensent que, bien que la plupart des modèles de traitement des phrases utilisent SPR, cette méthode présente certaines limites. Il peut y avoir des différences entre le comportement conscient de décision pour appuyer sur un bouton et le comportement inconscient des mouvements naturels des yeux. Comme dans l'expérience SPR le participant contrôle activement le temps et le rythme de la lecture, cette méthode a un temps de lecture plus lent que la lecture naturelle et pourrait avoir des biais dans les données. D'ailleurs, le comportement des mouvements oculaires peut avoir deux niveaux de données, à savoir le temps de lecture et la probabilité de régression, c'est-à-dire un retour en arrière spatial du regard. La méthode SPR, en revanche, ne contient que le temps de lecture.

Potentiels évoqués (angl. *Event-related potentials*)

Au chapitre 1 section 1.5.1, nous savons qu'il est possible de détecter les potentiels évoqués sur le traitement des phrases à travers des expériences en électroencéphalographie (EEG). L'avantage des expériences sur les potentiels évoqués est qu'elles peuvent identifier les effets de différents niveaux linguistiques sur le traitement des phrases. Par exemple, nous savons que le potentiel évoqué pertinent sur le traitement sémantique est la N400 (Kutas & Hillard, 1984 ; Dambacher et al., 2006), alors que si nous voulons étudier l'influence de la syntaxe dans le traitement des phrases, nous avons LAN (Friederici, 2011 ; 2002) et la P600 (Osterhout & Holcomb, 1992 ; Friederici et al., 2002 ; Kaan et al., 2000).

Cependant, deux obstacles nous empêchent d'évaluer systématiquement nos résultats de complexité syntaxique sur la base des potentiels évoqués. Premièrement, la plupart des études avec potentiels évoqués permettent de détecter des anomalies plutôt que d'observer des phrases ordinaires (Clifton & Staub, 2011). Deuxièmement, les études sur ce domaine ne sont pas aussi avancées qu'elles pourraient l'être en matière de complexité syntaxique. Clifton et Staub (2011)

soutiennent que les événements potentiels évoqués peuvent être utilisés comme une approche complémentaire à l'approche des mouvements oculaires.

Commentaires

Enfin, nous comparons ici toutes ces méthodes (Tableau 17) en fonction de deux propriétés. L'un est la faisabilité et l'autre la fiabilité.

Quant à ce qui est de la faisabilité, en général, nous considérons que les méthodes utilisant le nombre de mots correctement rapportés et le SPR sont les plus pratiques à réaliser par rapport aux deux autres. Puisque les mouvements oculaires et les potentiels évoqués nécessitent plus d'équipements expérimentaux comme nous l'avons dit, il faut plus de soutien financier dans ces approches, de plus elles impliquent plus de travail sur le traitement des données.

En ce qui concerne la fiabilité, les mouvements oculaires et les potentiels évoqués sont meilleurs que les deux autres approches. En effet, les mouvements oculaires peuvent avoir deux dimensions d'information, une pour le temps de lecture et une pour la probabilité de régression. Et les potentiels évoqués peuvent éviter l'influence du traitement sémantique sur le traitement des phrases.

Toutes les méthodes ont en fait leurs avantages et leurs limites. En conclusion, nous pensons qu'il se pourrait que les expériences sur les mouvements oculaires soient plus appropriées pour évaluer la complexité syntaxique. Comme ils reflètent également un processus de lecture plus naturel par rapport au SPR, et qu'ils peuvent traiter des phrases normales, ce qui n'est pas le cas pour des potentiels évoqués, leur approche pourrait être préférable.

Méthodes	Faisabilité	Fiabilité
Nombre de mots rapportés correctement	+++	+
SPR	+++	++
Mouvements oculaires	++	+++
ERP	++	+++

Tableau 17 Comparaison des méthodes d'évaluation

4.5 Conclusion

Dans ce chapitre, nous avons présenté des métriques du flux et nous avons pu voir leurs différentes hypothèses sur la nature de la complexité syntaxique. Ces métriques du flux concernent les arbres ordonnés, et elles permettent d'obtenir non seulement la complexité syntaxique de l'arbre entier, mais aussi la complexité syntaxique de chaque mot lors du traitement des phrases.

Nous venons de présenter dans ce chapitre diverses possibilités pour mesurer la complexité syntaxique en passant par la notion de flux ; cependant certaine d'entre elles ne nous conduisent pas immédiatement à une valeur numérique pour la complexité syntaxique, comme par exemple la métrique *C* qui comportent d'un paramètre *a* dans sa formulation (voir section 4.3.2.2). Nous tenons à préciser que la recherche des valeurs pour le paramètre est aussi une direction de travail qu'il nous resterait à explorer dans le futur.

Enfin dans la section 4.4, nous avons montré comment évaluer numériquement au travers de nos métriques une certaine forme de complexité syntaxique dans l'acception suivante : la difficulté dans le traitement de phrases existantes aussi bien à l'oral qu'à l'écrit. Par ailleurs nous ne nous attarderons pas sur les éventuelles distinctions entre les aspects *réception* ou *production*. Cependant il faudrait confirmer le lien de pertinence avec une réelle complexité au

travers d'expériences relevant d'autres domaines comme celui de la psycholinguistique et ceci reste donc à faire.

Notre prochain chapitre se concentre sur les études quantitatives des métriques que nous avons introduites : la taille du flux, le poids du flux, le poids combiné du flux, le flux potentiel et le flux requis. Nous disposons de suffisamment de données de corpus pour calculer ces métriques et pour vérifier les différentes hypothèses évoquées précédemment sur la complexité syntaxique en analysant les distributions des valeurs sur ces corpus. En outre, en prenant en compte la nature du corpus lui-même (voir chapitre 2), nous analyserons également les résultats selon différents types de corpus, ainsi que les résultats selon différentes typologies de langues.

5 Analyses quantitatives

5.1 Introduction

Dans ce chapitre, nous présentons nos études quantitatives des métriques du flux.

Tout d'abord, nous nous intéressons aux métriques du flux fondées sur un arbre syntaxique complet. Elles permettent d'examiner la complexité de la structure globalement. La section 5.2 présentera les analyses sur les distributions de la taille du flux et de la longueur de dépendance. La section 5.3 porte sur les études sur le poids du flux.

Ensuite, nous nous focaliserons sur une métrique du flux fondée sur un arbre partiel, qui est le flux potentiel projectif. L'arbre partiel est généralement la portion de structure qui précède la position inter-mot étudiée. Le flux potentiel projectif permet d'examiner la complexité du traitement de gauche à droite. Nous présenterons nos études du flux potentiel dans la section 5.4.

A la fin, nous présenterons les autres expériences dans la section 5.5. Ce sont des expériences qui soit n'en sont qu'à leur phase préliminaire, soit présentent une similarité dans les conclusions avec les expériences fondamentales que nous présentons dans les sections 5.2-5.4.

5.2 Taille du flux et longueur de dépendance

Dans cette section, nous présentons et prolongeons les travaux publiés par Kahane et Yan (2019) sur la taille du flux et la longueur de dépendance. Comme nous l'avons vu dans la section 4.3.1, la taille du flux et la longueur de la dépendance sont deux métriques différentes, qui induisent une même métrique sur les phrases, puisque la somme des tailles du flux est égale à la somme des longueurs des dépendances. Elles impliquent également deux hypothèses cognitives différentes sur le traitement des phrases. La taille du flux représente la quantité des

informations nécessaires qui sont chargées simultanément lors du traitement. La longueur de dépendance, d'autre part, représente la durée de stockage dans la mémoire de travail d'un mot qui attend son gouverneur ou son dépendant. Ainsi, les études quantitatives sur les distributions de ces deux métriques essayent de répondre à deux questions : premièrement il s'agit d'examiner si les valeurs de la taille du flux et de la longueur de dépendance sont contraintes. Si c'est le cas, il est important de trouver si elles sont toutes deux contraintes au même degré. Ensuite, il s'agit d'examiner leurs distributions dans les treebanks des différentes langues. L'objectif est de savoir si les résultats varient en fonction de la langue.

5.2.1 Distribution dans les treebanks UD

Les treebanks utilisés dans les expériences ci-dessous sont les treebanks UD en version 2.4, qui comprend 146 corpus et 83 langues (Nivre et al., 2019). En calculant la taille pour toutes les positions du flux et la longueur de toutes les dépendances dans ces treebanks, on constate, sans surprise car nos analyses exposées précédemment au chapitre 4 débouchaient sur de véritables propriétés générales, que la taille du flux et la longueur de la dépendance ont la même moyenne dans l'ensemble des treebanks UD (2,73 dans les deux cas) [46]. Il est nécessaire d'examiner aussi la distribution des deux ensembles de valeurs pour avoir une vision plus détaillée. Dans cette section, nous allons analyser la distribution de la taille du flux et de la longueur de dépendance dans l'ensemble des données UD. La section suivante se concentre sur la distribution des valeurs dans des langues différentes.

[46] Si les sommes sont égales, les moyennes le sont forcément.

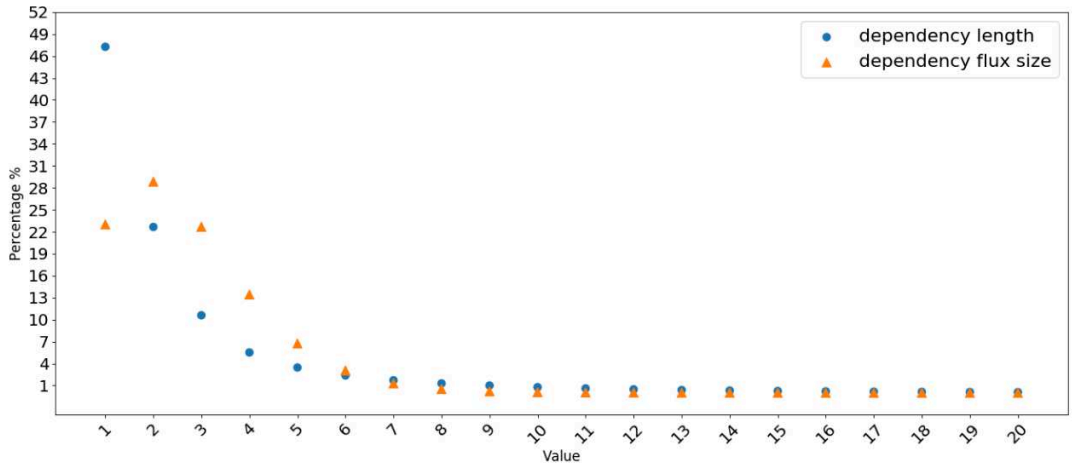

Figure 69. Répartition des longueurs de dépendance et des tailles du flux dans UD 2.4 (Kahane & Yan, 2019)

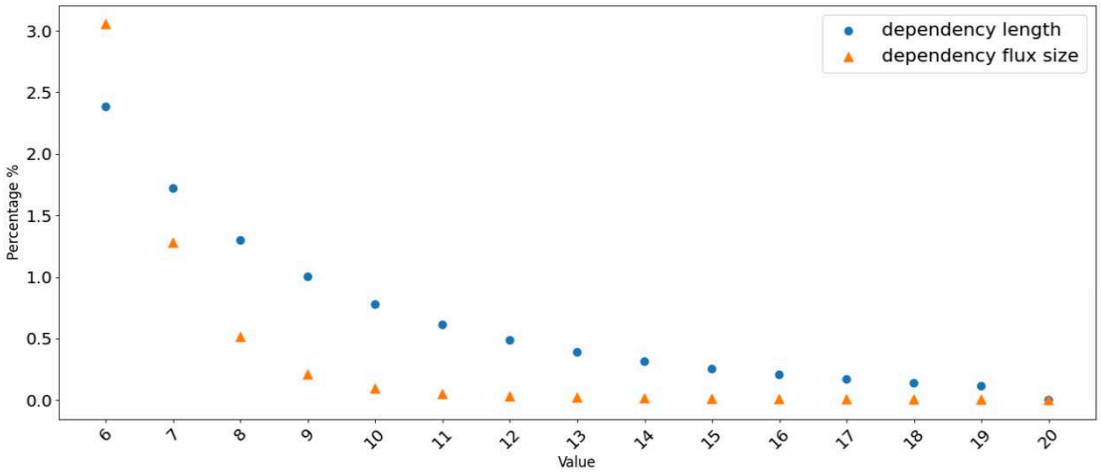

Figure 70. Répartition des valeurs entre 6 et 20 de longueur de dépendance et de taille du flux dans UD 2.4

La figure 69 et la figure 70 montrent la répartition des valeurs de longueur de dépendance et de taille du flux. Le triangle représente la fréquence de la taille du flux pour chaque valeur entre 1 et 20, et le point représente celui de la longueur de dépendance.

Pour la distribution des dépendances en fonction de leur longueur, on constate que 47,26% des dépendances ont une longueur égale à 1, puis la fréquence de la longueur diminue rapidement avec l'augmentation de la valeur (voir aussi le tableau 18 pour les détails).

En ce qui concerne la distribution des positions du flux en fonction de leur taille, les positions du flux ayant une taille égale à 2 sont plus nombreux que celles de taille 1 (28,84% contre

22,66%). Après la valeur 2, les positions du flux sont progressivement moins nombreuses.

Si on regarde les deux courbes, on observe une diminution plus lente pour la courbe de la taille du flux au début par rapport à celle de la longueur de dépendance, puis la diminution s'accélère. Les deux courbes se croisent pour la valeur 7. Après cette valeur, le pourcentage de la taille du flux est inférieur à celui de la longueur de dépendance.

n	% LD = n	% LD ≤ n	% TF = n	% TF ≤n
1	47,2591	47,2591	22,9980	22,9980
2	22,6573	69,9164	28,8418	51,8398
3	10,5948	80,5112	22,6748	74,5146
4	5,5159	86,0271	13,4303	87,9449
5	3,4615	89,4886	6,7401	94,6851
6	2,3822	91,8709	3,0540	97,7391
7	1,7192	93,5901	1,2787	99,0178
8	1,2969	94,8870	0,5126	99,5304
9	1,0014	95,8884	0,2087	99,7391
10	0,7763	96,6647	0,0935	99,8326
11	0,6111	97,2758	0,0488	99,8814
12	0,4856	97,7614	0,0299	99,9113
13	0,3882	98,1495	0,0203	99,9316
14	0,3126	98,4621	0,0142	99,9457
15	0,2522	98,7143	0,0100	99,9557
16	0,2054	98,9196	0,0073	99,9630
17	0,1682	99,0878	0,0056	99,9686
18	0,1370	99,2248	0,0045	99,9731
19	0,1130	99,3378	0,0036	99,9767
20	0,0937	99,4315	0,0032	99,9799

21	0,0782	99,5097	0,0029	99,9827
22	0,0641	99,5738	0,0025	99,9852
23	0,0541	99,6280	0,0020	99,9872
24	0,0454	99,6733	0,0017	99,9890
25	0,0388	99,7121	0,0015	99,9904

Tableau 18. Distribution des longueurs de dépendance et des tailles du flux dans UD 2.4 (LD : longueur de dépendance ; TF: taille du flux ; n : valeur de longueur ou de taille) (Kahane & Yan, 2019)

Nous appelons ici **la valeur contrainte** la valeur pour laquelle 99,9% des tailles ou des longueurs sont égales ou inférieures. Le tableau 18 montre lui aussi les fréquences pour les longueurs et les tailles ayant une valeur ≤ n (n est entre 0 et 25).

Nous pouvons constater que sur nos corpus, 99,9% des positions inter-mot ont une taille de flux inférieure ou égale à 12. La valeur contrainte pour la taille du flux est de 12. Alors que pour la longueur de dépendance, il faut atteindre la valeur 36 pour avoir plus de 99,9% des dépendances. Ainsi, sa valeur contrainte est de 36.

5.2.2 Distribution dans des treebanks de différentes langues

En examinant les treebanks séparément, on a obtenu des résultats similaires à ce que nous avons trouvé dans la section précédente globalement. Le tableau 19 montre les distributions des valeurs de 47 treebanks contenant plus de 100 000 positions de flux.

Tout d'abord on peut constater les valeurs contraintes pour la taille du flux et pour la longueur de dépendance dans chaque treebank. Comme le tableau 19 le montre, cette valeur pour la longueur de dépendance est entre 17 (*UD_Finnish-FTB*) et 75 (*UD_Arabic-NYUAD, UD_Arabic-PADT*), et la moyenne pour ces 47 treebanks est de 32,64. En ce qui concerne la taille du flux, elle se trouve entre 8 (*UD_Bulgarian-BTB*, *UD_Czech-FicTree*, *UD_Korean-Kaist, UD_Old_French-SRCMF*, *UD_Polish-PDB*, *UD_Romanian-RRT*) et 32 (*UD_Ancient_Greek-Perseus*), et la moyenne pour ces 47 treebanks est de 11,30. Comme dans le cas des résultats

de l'ensemble de données UD, la taille du flux est beaucoup plus contrainte que la longueur de dépendance dans les treebanks.

Croisements des deux courbes de distribution et type des langues

Pour comparer la distribution de la taille du flux avec celle de la longueur de dépendance dans les treebanks, on peut observer les croisements de ces deux courbes dans le graphique (du type Figure 69) pour chaque treebank. Le croisement signifie que le sens de la comparaison entre les deux distributions commence à changer. Pour la distribution de la longueur de dépendance et celle de la taille du flux, il s'agit tout d'abord d'un croisement où **le pourcentage de la taille commence par être supérieur à celui de la longueur de dépendance,** on le représente ici **C↑;** Ensuite, le deuxième croisement indique que **le pourcentage de la taille descend plus vite que celui de la longueur de dépendance, et que donc il commence à être inférieur au dernier** (**C↓**)**.** Comme il est assez volumineux de faire paraître tous les graphiques, le tableau 19 résume les informations sur tous les croisements pour les treebanks.

Premièrement, les courbes pour la longueur de dépendance et la taille du flux se croisent toujours (voir les données de la colonne C↑ dans le tableau 19) à la valeur 2 : la fréquence de la taille du flux égale à 1 est toujours inférieure à celle de la longueur de dépendance, tandis que la fréquence de la taille du flux égale à 2 est supérieure à celle de la longueur de dépendance.

Ensuite, la fréquence pour la taille diminue très rapidement après la valeur 2, et un second croisement (C↓ dans tableau 19) se produit entre la valeur 5 (*UD_Finish-FTB*), et la valeur 8 (dans 9 treebanks: *UD_Urdu-UDTB*, *UD_Persian-Seraji*, *UD_Hindi-HDTB*, *UD_German-HDT*, *UD_German-GSD*, *UD_Dutch-Alpino*, *UD_Chinese-GSD*, *UD_Arabic-PADT* et *UD_Japanese-BCCWJ*). Après cette valeur, la fréquence de la taille du flux est inférieure à celle de la longueur de dépendance, et la première diminue beaucoup plus vite que la dernière.

Ce phénomène est accentué pour les résultats des fréquences cumulées[47]. Le croisement pour

[47] La fréquence cumulée pour une valeur *n* est la fréquence de la taille ou de la longueur inférieure ou égale à *n*.

les deux courbes de fréquences cumulées C↓FC (voir le tableau *19*) se fait entre les valeurs 3 (dans 4 treebanks : *UD_Estonian-EDT*, *UD_Finnish-FTB*, *UD_Finnish-TDT*, et *UD_Polish-PDB*) et 5 (dans les mêmes 9 treebanks que précédemment). Dans les treebanks avec le croisement C↓FC à la valeur 5, la plupart sont des langues à tête finale comme le japonais (*UD_Japanese-BCCWJ*), l'allemand (*UD_German-GSD*, *UD_German-HDT*), le néerlandais (*UD_Dutch-Alpino*) et le persan (*UD_Persian-Seraji*), ainsi que le chinois (*UD_Chinese-GSD*). Une exception est l'arabe (*UD_Arabic-PADT*), qui est elle une langue à tête initiale.

Dans les treebanks avec le croisement C↓FC à la valeur 3, il s'agit de langues à tête initiale comme le finnois (*UD_Finnish-FTB* et *UD_Finnish-TDT*), le polonais (*UD_Polish-PDB*) et l'estonien (*UD_Estonian-EDT*).

n	**% LD ≤ n**	**% TD ≤ n**	**C↑**	**C↓**	**C↓ FC**
UD_Ancient_Greek-Perseus*	34	32	2	7	4
UD_Ancient_Greek-PROIEL	34	14	2	7	4
UD_Arabic-NYUAD	75	20	2	7	4
UD_Arabic-PADT	75	17	2	8	5
UD_Bulgarian-BTB	21	8	2	6	4
UD_Catalan-AnCora	38	10	2	7	4
UD_Chinese-GSD	40	12	2	8	5
UD_Croatian-SET	27	9	2	7	4
UD_Czech-CAC	32	13	2	6	4
UD_Czech-FicTree	23	8	2	6	4
UD_Czech-PDT*	25	9	2	6	4
UD_Dutch-Alpino	27	10	2	8	5
UD_English-EWT	31	10	2	7	4

UD_Estonian-EDT	22	9	2	6	3
UD_Finnish-FTB	17	9	2	5	3
UD_Finnish-TDT	30	15	2	6	3
UD_French-FTB	41	11	2	7	4
UD_French-GSD	31	11	2	6	4
UD_Galician-CTG	28	9	2	7	4
UD_German-GSD	29	11	2	8	5
UD_German-HDT	26	10	2	8	5
UD_Hebrew-HTB	37	11	2	6	4
UD_Hindi-HDTB	33	11	2	8	5
UD_Italian-ISDT	35	13	2	6	4
UD_Italian-PoSTWITA	22	10	2	7	4
UD_Italian-VIT	39	9	2	7	4
UD_Japanese-BCCWJ	60	20	2	8	5
UD_Japanese-GSD	44	9	2	7	4
UD_Korean-Kaist	19	8	2	6	4
UD_Latin-ITTB	27	9	2	6	4
UD_Latin-PROIEL	35	14	2	7	4
UD_Latvian-LVTB	27	9	2	6	4
UD_Norwegian-Bokmaal	24	9	2	6	4
UD_Norwegian-Nynorsk	26	10	2	6	4
UD_Old_French-SRCMF	23	8	2	6	4
UD_Old_Russian-TOROT	29	16	2	6	4

UD_Persian-Seraji	40	11	2	8	5
UD_Polish-PDB	24	8	2	6	3
UD_Portuguese-Bosque	34	9	2	7	4
UD_Portuguese-GSD	33	10	2	6	4
UD_Romanian-Nonstandard	27	9	2	6	4
UD_Romanian-RRT	28	8	2	6	4
UD_Russian-SynTagRus	26	9	2	6	4
UD_Slovenian-SSJ	25	9	2	7	4
UD_Spanish-AnCora	34	9	2	7	4
UD_Spanish-GSD	34	9	2	6	4
UD_Urdu-UDTB	43	17	2	8	5

Tableau 19. Répartition des longueurs de dépendance (LD) et des tailles de flux (TF) dans 47 treebanks de UD 2.4 (C↓ FC : croisement C↓ de deux courbes de fréquences cumulées) * Il y a d'autres croisements, mais ils sont après la valeur 40 qui sont assez peu représentatifs.

5.2.3 Commentaires

Les études quantitatives dans l'ensemble des treebanks UD montrent que ce sont les faibles valeurs pour les deux métriques qui sont les plus fréquentes. En d'autres termes la fréquence des grandes tailles ou des grandes longueurs est faible. Du point de vue cognitif, nous savons déjà que la longueur de dépendance est liée à la durée de stockage d'un élément dans la mémoire de travail. La minimisation de longueur de dépendance résulte de la contrainte suivante : le gouverneur et le dépendant ont tendance à se rapprocher pour éviter de les garder trop longtemps dans la mémoire de travail. La taille du flux permet d'interpréter la minimisation de longueur de dépendance autrement (puisque comme nous l'avons déjà dit la longueur moyenne de dépendance est égale à la taille moyenne du flux). Comme la taille du flux représente la quantité d'informations concomitantes à traiter au cours de chaque position

inter-mot, cela signifie que l'interlocuteur a aussi tendance à minimiser les informations concomitantes pendant le traitement des phrases.

Il est important de savoir qu'avoir une tendance à être minimisé n'est pas la même chose qu'être borné. La borne en valeur d'une métrique est une valeur impossible à dépasser. Cependant, d'un point de vue théorique, on peut augmenter la taille du flux en augmentant le nombre d'éléments conjonctifs, de même pour des modifieurs de verbe, sans limite en UD. En SUD, les éléments conjonctifs ne seront pas en bouquet, donc l'augmentation de la taille du flux ne se fait pas de la même façon en ce qui concerne les éléments conjonctifs, et la taille du flux pourrait être plus ou moins bornée sauf à considérer les modifieurs. En ce qui concerne la longueur de dépendance, nous pensons qu'elle n'est pas bornée. On peut avoir par ailleurs une très longue dépendance dans une phrase assez longue, mais nous ne pensons pas que seule l'augmentation de longueur pour une dépendance provoque l'échec du traitement de la phrase (voir chapitre 1 au niveau de l'auto-enchâssement).

En regardant leurs distributions pour 99,9% des positions, nous avons constaté que les valeurs de ces deux métriques ne sont pas contraintes au même degré. Comparée avec la longueur de dépendance, la taille diminue beaucoup plus vite en pourcentage. On constate que 99,9% de positions ont une taille inférieure ou égale à 12, et que pour la longueur de dépendance cette valeur est de 36. Ainsi, la taille du flux est plus contrainte que la longueur de dépendance.

De plus, en examinant chaque treebank de langues différentes, on constate que la contrainte sur la taille du flux est plus stable que la longueur de dépendance. De fait, la longueur de la dépendance varie plus fortement et est plus sensible aux langues ou aux genres de corpus (Liu, 2008 ; Jiang et Liu, 2015).

En outre, la distribution des tailles du flux est également liée au type de langue. Les résultats présentés précédemment (Tableau 18) nous ont fait remarquer que les tailles du flux sont généralement plus importantes dans les langues à tête finale que dans les langues à tête initiale. Cela indique également que la taille du flux serait plus contrainte dans les langues à tête initiale. Toutefois, cette hypothèse nécessite des travaux ultérieurs pour être confirmée. Nous

reviendrons sur cette discussion lorsque nous comparerons ces résultats avec ceux du flux potentiel et du flux requis dans les sections 5.4 et 5.5.3.

5.3 Poids du flux

Dans cette section nous présentons nos études quantitatives sur le poids du flux. Le poids du flux est discuté dans la section 4.3.2 du chapitre 4, nous rappelons qu'il est le nombre maximal de dépendances disjointes. Cette métrique permet d'étudier le niveau d'auto-enchâssement des constructions.

Cette section est divisée en trois parties principales :

- La première partie (section 5.3.1) concerne la problématique de la limitation universelle du poids du flux dans toutes les langues. Nous avons expliqué précédemment que le poids du flux représente le niveau d'auto-enchâssement des constructions. Le niveau d'auto-enchâssement est limité chez l'homme en raison de contraintes cognitives selon plusieurs études (Miller & Chomsky, 1965 ; Bever, 1970 ; Lewis, 1996 ; Gibson, 2000), les études quantitatives tentent donc dans un premier temps de vérifier cette hypothèse.
- La deuxième partie (section 5.3.2) examine précisément la distribution des poids dans les corpus français de différents genres. Comme il peut y avoir des différences sur la complexité syntaxique entre les corpus parlés et écrits, nous examinerons la distribution des poids dans les corpus parlés et écrits.
- La troisième partie (section 5.3.3) concerne la distribution des poids de flux dans des arbres ayant différentes granularités (voir la section 2.4.2 du chapitre 2). En traitant des arbres syntaxiques dans les treebanks, il est possible d'obtenir des arbres syntaxiques qui ne prennent en compte que les dépendances mettant en jeu des mots pleins en agrégeant mots pleins et mots fonctionnels. Utiliser cette granularité d'analyse réduit le déséquilibre des annotations entre les langues (ce qui rend les langues plus comparables). Les résultats des poids de flux deviennent plus homogènes. On pourra,

dans un travail subséquent, avoir une démarche analogue pour les autres métriques, celle de la longueur de dépendance et celle de la taille du flux.

5.3.1 Poids du flux dans les treebanks UD

Les poids du flux ont été étudiés dans les travaux de Kahane, Yan et Botalla (2017) et de Yan (2017). Les calculs ont été effectués à partir des treebanks UD en version 2.0, qui comprenaient 70 treebanks en 50 langues. Ces travaux montrent que le poids maximum est entre 3 (*UD_Sanskrit*) et 7 (UD_*Czech-CLTT*) dans les treebanks[48].

Le tableau 20 montre la répartition des poids du flux de 1 à 7 pour chaque treebank. La dernière ligne montre la distribution des poids de flux pour l'ensemble des treebanks UD. En outre, des informations sur la taille de chaque treebank sont également fournies dans le tableau, car différentes tailles de treebanks peuvent avoir un impact sur les résultats finaux. Cela est particulièrement vrai pour les petits treebanks.

Résultats pour l'ensemble des données d'UD

Pour l'ensemble des données d'UD, la plupart des poids ont tendance à être inférieurs à 4 : 62,15 % des poids sont égaux à 1, 94,9% sont inférieurs ou égaux à 2, et 99,61% sont inférieurs ou égaux à 3. Seulement 0,36% des poids sont égaux à 4, et 0,022 % des poids sont supérieurs à 4.

Résultats par treebank

Quels que soient les treebanks UD, les positions inter-mot ayant un poids égal ou supérieur à 5 sont extrêmement rares. La majorité des treebanks contiennent très peu de positions avec des valeurs de poids égales à 5 : entre 0% et 0,34% (*UD_Uyghur*). De plus, la majorité des

[48] Nous constatons qu'il s'agit d'un cas exceptionnel dans *UD_Czech-CLTT* ayant un poids à 7 et que cela n'est sans doute pas généralisable.

treebanks n'ont presque aucune position dans le flux avec un poids de 6. Les cas plutôt exceptionnels sont pour *UD_Czech-CLTT* qui a 0,048% de positions avec un poids de 6, *UD_Hungarian* qui a 0,017% et *UD_Chinese* qui a 0,013%.

Ensuite, il existe des différences dans la répartition du poids pour les treebanks. Comme la taille des treebanks peut affecter les résultats, on n'a pris en compte que ceux qui ont au moins 10 000 positions. Certains treebanks ont des poids relativement faibles, par exemple, *UD_Finnish-FTB*, *UD_Polish* et *UD_Slovak* ont 80% des poids égaux à 1. Cependant, pour les treebanks en arabe, coréen et chinois, les poids sont généralement plus importants, avec plus de 10% des poids égaux à 3.

Enfin, la plupart des treebanks d'une même langue présentent des résultats similaires. Cependant, il y a aussi des exceptions, par exemples le cas du tchèque : *UD_Czech-CLTT* ont des poids beaucoup plus importants que *UD_Czech-CAC* et *UD_Czech*. Nous pensons qu'il y a deux raisons principales, l'une est que les treebanks peuvent avoir été développés par des équipes différentes, et l'analyse syntaxique entre elles peut être différente. L'autre est que cela peut être dû au genre des textes. Le *UD_Czech-CLTT* est composé de textes juridiques ; le *UD_Czech* est composé de textes journalistiques et encyclopédiques ; le *UD_Czech-CAC* est composé de textes juridiques, médicaux, et journalistiques et des commentaires. Comme le CLTT ne contient que des textes juridiques rédigés en langage sophistiqué, cela pourrait créer des poids plus importants que les textes d'autres genres.

Corpus	Trees	1	2	3	4	5	6	7
UD_Ancient_Greek	12613	57,7713%	35,7922%	5,9551%	0,4521%	0,0249%	0,0045%	0
UD_Ancient_Greek-PROIEL	15865	57,8084%	35,9672%	5,7419%	0,4633%	0,0187%	0,0005%	0
UD_Arabic	6984	47,1524%	41,1035%	10,5967%	1,0961%	0,0513%	0	0
UD_Arabic-NYUAD	19738	47,1593%	40,8588%	10,6710%	1,2305%	0,0778%	0,0026%	0
UD_Basque	7194	67,8464%	28,2782%	3,6570%	0,2122%	0,0061%	0	0
UD_Belarusian	333	62,4331%	31,6951%	5,3690%	0,5028%	0	0	0
UD_Bulgarian	10022	73,8611%	24,2038%	1,8737%	0,0598%	0,0017%	0	0
UD_Catalan	14832	57,8165%	36,5795%	5,2978%	0,2958%	0,0102%	0,0002%	0
UD_Chinese	4497	49,7275%	37,3540%	10,9053%	1,7849%	0,2156%	0,0127%	0
UD_Coptic	320	61,7610%	33,7425%	4,3712%	0,1253%	0	0	0
UD_Croatian	8289	64,4593%	31,7047%	3,6612%	0,1682%	0,0066%	0	0
UD_Czech	77765	66,7821%	29,6130%	3,4235%	0,1748%	0,0065%	0,0001%	0
UD_Czech-CAC	24081	65,1560%	30,8173%	3,8019%	0,2183%	0,0062%	0,0002%	0
UD_Czech-CLTT	814	42,7840%	41,7363%	12,1886%	2,7072%	0,5319%	0,0480%	0,0040%
UD_Danish	4947	69,1298%	27,8483%	2,9037%	0,1182%	0	0	0
UD_Dutch	13050	63,4139%	31,1239%	4,9995%	0,4395%	0,0232%	0	0
UD_Dutch-LassySmall	6841	70,3049%	27,0392%	2,5037%	0,1522%	0	0	0

UD_English	14545	68,4521%	28,0830%	3,2899%	0,1696%	0,0045%	0,0010%	0
UD_English-LinES	3650	68,3978%	28,1629%	3,2005%	0,2288%	0,0100%	0	0
UD_English-ParTUT	1590	64,9917%	31,0885%	3,7673%	0,1497%	0,0029%	0	0
UD_Estonian	3172	77,2578%	20,3719%	2,2120%	0,1475%	0,0108%	0	0
UD_Finnish	13581	72,5965%	23,7918%	3,2765%	0,2981%	0,0345%	0,0026%	0
UD_Finnish-FTB	16856	82,7698%	15,9091%	1,2182%	0,0985%	0,0044%	0	0
UD_French	16031	64,3145%	32,2325%	3,2669%	0,1741%	0,0120%	0	0
UD_French-ParTUT	620	60,6591%	34,9931%	4,0532%	0,2646%	0,0301%	0	0
UD_French-Sequoia	2643	61,2494%	33,7609%	4,6907%	0,2863%	0,0127%	0	0
UD_Galician	3139	62,7834%	33,7314%	3,3389%	0,1433%	0,0029%	0	0
UD_Galician-TreeGal	600	61,9842%	33,7488%	4,0206%	0,2464%	0	0	0
UD_German	14917	59,6002%	35,5981%	4,4572%	0,3294%	0,0135%	0,0016%	0
UD_Gothic	4372	65,8269%	30,5083%	3,4588%	0,2061%	0	0	0
UD_Greek	2065	62,4918%	33,7292%	3,5903%	0,1866%	0,0021%	0	0
UD_Hebrew	5725	58,0375%	36,5604%	5,1693%	0,2299%	0,0029%	0	0
UD_Hindi	14963	49,0185%	44,7635%	5,6525%	0,5365%	0,0262%	0,0028%	0
UD_Hungarian	1351	56,0360%	35,2358%	7,5811%	0,9872%	0,1425%	0,0174%	0
UD_Indonesian	5036	64,8235%	31,3805%	3,6083%	0,1816%	0,0060%	0	0
UD_Irish	566	53,1129%	38,5677%	7,4709%	0,8169%	0,0317%	0	0
UD_Italian	13402	64,9254%	31,5456%	3,3684%	0,1554%	0,0051%	0	0

UD_Italian-ParTUT	1590	61,5631%	34,4833%	3,7806%	0,1705%	0,0025%	0	0
UD_Japanese	7675	50,9777%	42,8164%	6,0326%	0,1720%	0,0013%	0	0
UD_Kazakh	31	55,2688%	37,4194%	6,8817%	0,4301%	0	0	0
UD_Korean	5350	51,2951%	37,0994%	10,2358%	1,2832%	0,0865%	0	0
UD_Latin	1334	56,3420%	35,8951%	7,0072%	0,6910%	0,0646%	0	0
UD_Latin-ITTB	16508	60,4397%	33,2464%	5,8513%	0,4457%	0,0166%	0,0004%	0
UD_Latin-PROIEL	15324	61,3042%	31,4055%	6,2665%	0,9210%	0,1006%	0,0021%	0
UD_Latvian	3054	67,2067%	27,5104%	4,7455%	0,4800%	0,0522%	0,0052%	0
UD_Lithuanian	263	66,7570%	28,9677%	4,0667%	0,2086%	0	0	0
UD_Norwegian-Bokmaal	18106	72,7316%	24,9549%	2,2322%	0,0809%	0,0004%	0	0
UD_Norwegian-Nynorsk	16064	70,7325%	26,6657%	2,4955%	0,1043%	0,0016%	0,0004%	0
UD_Old_Church_Slavonic	5196	69,3145%	27,4069%	3,1463%	0,1276%	0,0047%	0	0
UD_Persian	5397	45,7475%	45,1357%	8,5169%	0,5881%	0,0111%	0,0008%	0
UD_Polish	7127	82,4621%	16,9600%	0,5744%	0,0035%	0	0	0
UD_Portuguese	8891	61,6894%	33,9385%	4,1555%	0,2115%	0,0050%	0	0
UD_Portuguese-BR	10874	60,0570%	35,5035%	4,2351%	0,1987%	0,0057%	0	0
UD_Romanian	8795	64,4173%	31,6192%	3,7420%	0,2171%	0,0038%	0,0005%	0
UD_Russian	4429	66,7603%	29,7524%	3,2753%	0,2081%	0,0038%	0	0
UD_Russian-SynTagRus	55398	67,3001%	29,0431%	3,4168%	0,2263%	0,0125%	0,0011%	0
UD_Sanskrit	190	71,9488%	27,0669%	0,9843%	0	0	0	0

UD_Slovak	9543	83,0950%	16,2359%	0,6595%	0,0096%	0	0	0
UD_Slovenian	7212	72,1230%	25,4867%	2,3071%	0,0796%	0,0036%	0	0
UD_Slovenian-SST	2137	70,0888%	26,5897%	3,1676%	0,1539%	0	0	0
UD_Spanish	15587	61,5361%	34,7461%	3,5564%	0,1539%	0,0075%	0	0
UD_Spanish-AnCora	15959	58,1968%	36,4534%	5,0973%	0,2481%	0,0043%	0	0
UD_Swedish	4807	70,7716%	26,6201%	2,4991%	0,1032%	0,0060%	0	0
UD_Swedish-LinES	3650	69,3245%	27,5105%	3,0290%	0,1342%	0,0017%	0	0
UD_Tamil	600	58,1370%	36,1832%	5,2254%	0,4544%	0	0	0
UD_Turkish	4660	60,7080%	31,6928%	6,6877%	0,8495%	0,0594%	0,0026%	0
UD_Ukrainian	863	74,8420%	23,6048%	1,4900%	0,0632%	0	0	0
UD_Urdu	4595	44,9431%	45,0896%	8,8165%	1,0376%	0,1131%	0	0
UD_Uyghur	100	43,8741%	42,1629%	11,4990%	2,1218%	0,3422%	0	0
UD_Vietnamese	2200	76,0973%	22,7062%	1,1816%	0,0148%	0	0	0
Total	630518	62,1538%	32,7521%	4,7116%	0,3608%	0,0208%	0,0009%	0,0000092%

Tableau 20. Poids du flux pour UD 2.0 (Kahane, Yan & Botalla, 2017 ; Yan, 2017)

Les exemples

Nous avons extrait des exemples dans les trois langues que nous connaissons le mieux, le français, l'anglais et le chinois. Ces arbres extraits sont ceux qui ont les plus grands poids (supérieurs ou égaux à 5)[49].

Trois observations principales ont été tirées de ces exemples :

- Tout d'abord, nous avons constaté que les arbres extraits sont des phrases très longues. Et ils contiennent souvent les dépendances en *parataxis*[50] ou *conj* (conjonction de coordination) qui rendent les poids plus importants. Mais cela ne veut pas dire que ces phrases sont effectivement plus compliquées à traiter, car les relations *parataxis* et *conj* lient les éléments qui ne forment pas de structure hiérarchique mais des structures parallèles.
- La deuxième remarque est que, dans les dépendances disjointes, les dépendances fonctionnelles comme *det* et *case* en font souvent partie. Pourtant elles ne sont pas si compliquées à traiter par rapport aux dépendances entre les mots pleins, comme *nmod* ou *nsubj* etc. Les sections qui suivent considéreront les arbres avec différentes granularités d'analyse (voir section 5.3.3 et section 2.4.2).
- Le dernier point est que dans les treebanks UD version 2.0, il s'agit aussi des poids importants qui sont causés par les annotations ayant des erreurs.

Nous allons monter des exemples ici qui concernent ces trois points pour le français, l'anglais

---

[49] Voir tous les arbres extraits sur https://github.com/chunxiaoyan/results_experiments_phd

[50] *"The parataxis relation (from Greek for "place side by side") is a relation between a word (often the main predicate of a sentence) and other elements, such as a sentential parenthetical or a clause after a ":" or a ";", placed side by side without any explicit coordination, subordination, or argument relation with the head word. Parataxis is a discourse-like equivalent of coordination, and so usually obeys an iconic ordering. Hence it is normal for the first part of a sentence to be the head and the second part to be the parataxis dependent, regardless of the headedness properties of the language. But things do get more complicated, such as cases of parentheticals, which appear medially." (https://universaldependencies.org/u/dep/parataxis.html)*

et le chinois.

Pour le français (poids supérieur ou égal à 5) :

- UD_French-Séquoia : 3 phrases (longueur de phrase : 52, 87 et 106)
- UD_French-Partut : 1 phrase (longueur de phrase : 140)
- UD_French(-GSD) : 8 phrases (longueur de phrase : 45, 108, 49,293, 82, 36, 72 et 59)

La figure 71 montre un arbre avec *parataxis* et la figure 72 montre un arbre avec *conj*.

Nous observons que l'existence de dépendance *parataxis* fait augmenter le poids du flux pour beaucoup de positions inter-mot. Le poids le plus important se trouve entre *à* et *le*[51] marqué par une ligne verticale pour l'exemple de la figure 71 :

- Niveau poids 1 : {à -*fixed*-> le, à-*fixed*-> moins}
- Niveau Poids 2 : {à <-*advmod*- mg, d' <-*case*- mg}
- Niveau poids 3 : {apport -*nmod*-> mg, apport -*nmod*-> fois}
- Niveau poids 4 : {administrer -*obl*-> jour}
- Niveau poids 5 : {conseillé -*parataxis*-> voir}

En ce qui concerne l'arbre dans la figure 72, ce sont les dépendances *conj* qui font augmenter le poids du flux pour beaucoup de positions inter-mot. Un des poids le plus important se trouve entre entre *de* et *reférence* marqué par une ligne verticale :

- Niveau poids 1 : {de <-*case*- référence}
- Niveau Poids 2 : {code -*nmod*-> référence, code -*conj*-> indéntité}
- Niveau poids 3 : {nom -*conj*-> code}
- Niveau poids 4 : {comptait -*conj*-> reprenais}
- Niveau poids 5 : {première -*conj*-> troisième}

[51] Les amalgames sont défaits pendant la tokenisation. Par exemples : au = à le, du = de le, des = de les

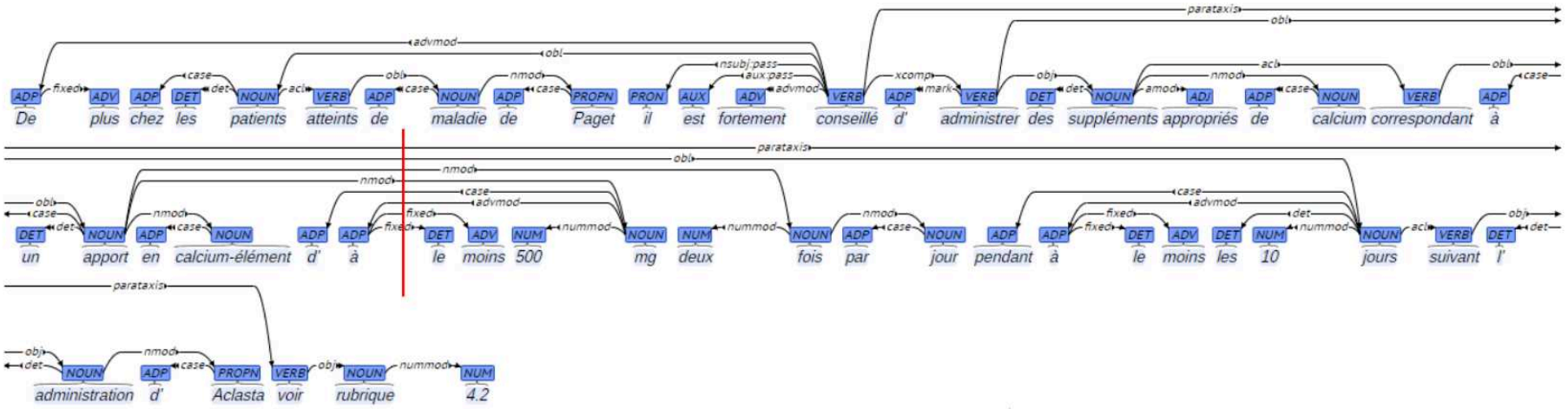

Figure 71 Exemple (UD_French-Séquoia, emea-fr-test_00119)

« *De plus, chez les patients atteints de maladie de Paget, il est fortement conseillé d'administrer des suppléments appropriés de calcium correspondant à un apport en calcium-élément d'au moins 500 mg deux fois par jour pendant au moins les 10 jours suivant l'administration d'Aclasta (voir rubrique 4.2).* »

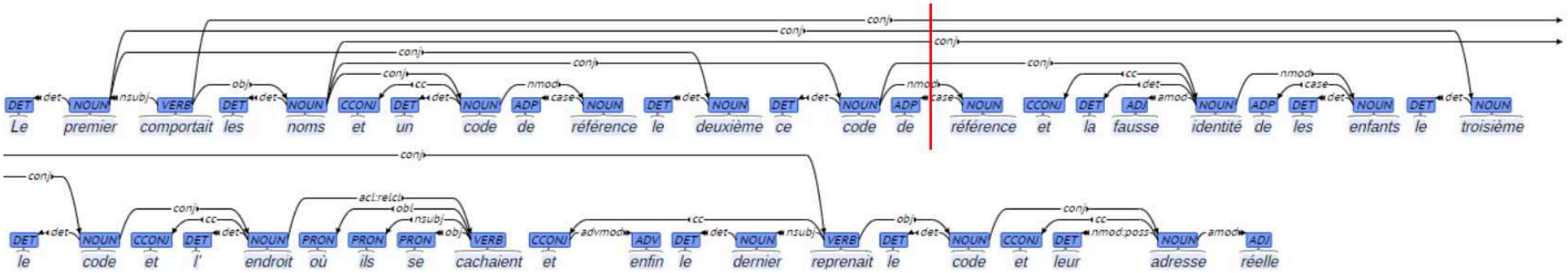

Figure 72 Exemple (UD_French(GSD), fr-ud-train_02640)

« *Le premier comportait les noms et un code de référence, le deuxième, ce code de référence et la fausse identité des enfants, le troisième, le code et l'endroit où ils se cachaient et enfin, le dernier reprenait, le code et leur adresse réelle.* »

Pour l'anglais (poids supérieur ou égal à 5) :

- UD_English-Partut : 1 phrase (longueur de phrase : 111)
- UD_English-LinES : 3 phrases (longueur de phrase : 55, 79 et 71)
- UD_English(-GSD) : 7 phrases (longueur de phrase : 40, 58, 53, 34, 43, 32, et 31)

Similaire au cas du français, ce sont des dépendances *conj* ou *parataxis* qui augmentent les poids. La figure 73 et la figure 74 montrent deux exemples.

On a marqué une ligne verticale sur une position ayant le poids à 5, qui est la position (*doctor*, *'s*) dans l'arbre de la figure 73 :

- Niveau poids 1 : {doctor -*case*-> 's}
- Niveau poids 2 : {doctor <-*nmod:poss*- offices, some <-*det*- offices}
- Niveau poids 3 : {way – *acl:relcl*-> work}
- Niveau poids 4 : {way <- *nsubj*- absurd, that <-*mark*- absurd}
- Niveau poids 5 : {feel - *ccomp*-> absurd, feel - *parataxis*-> happy}

En ce qui concerne l'arbre de la figure 74, on a marqué une ligne verticale sur une position du poids à 5, qui est la position (*of*, *6*) :

- Niveau poids 1 : {of <-*case*- 6, out <-*case*-6}
- Niveau poids 2 : {5 -*nmod*-> 6, 5 <-*compound*- monthly}
- Niveau poids 3 : {for <-*case*- premiums}
- Niveau poids 4 : {payment -*nmod*-> premiums, payment -*nmod*-> policy}
- Niveau poids 5 : {took -*conj*-> cancelled}

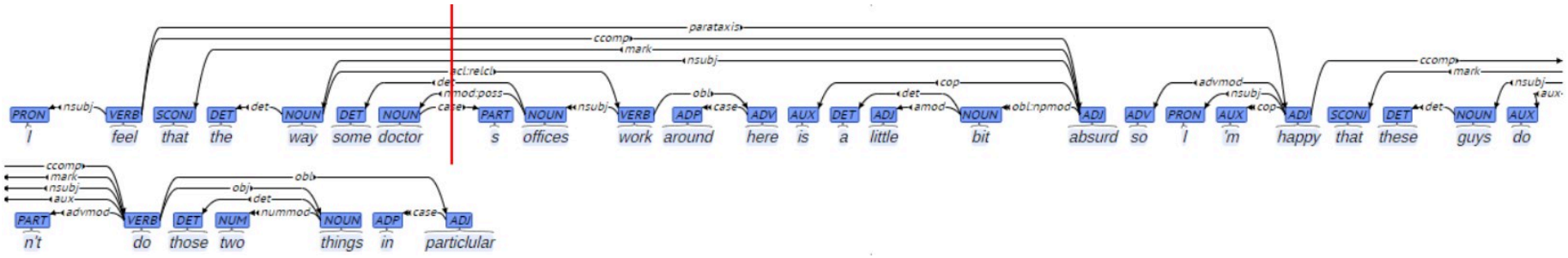

Figure 73 Exemple (UD_English-EWT, reviews-220214-0004)

"*I feel that the way some doctors offices work around here is a little bit absurd so I'm happy that these guys don't do those two things in particular.*"

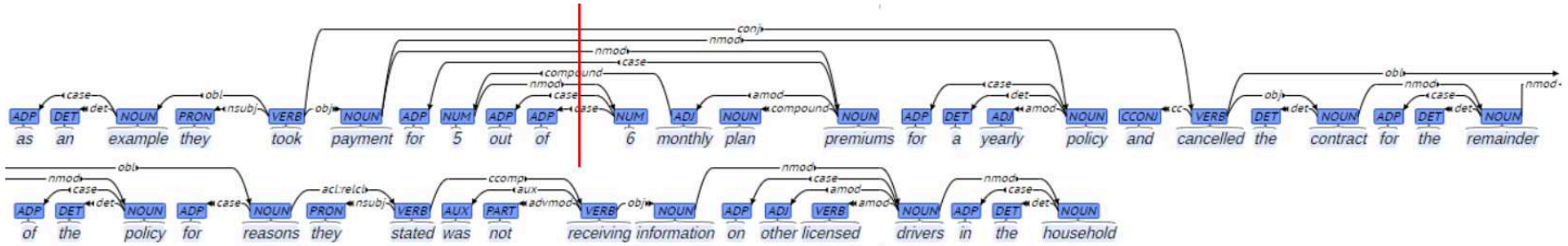

Figure 74 Exemple (UD_English-EWT, reviews-217359-0006)

"*as an example they took payment for 5 out of 6 monthly plan premiums for a yearly policy and cancelled the contract for the remainder of the policy for reasons they stated was not receiving information on other licensed drivers in the household?*"

Le cas du chinois est plus compliqué, il y a beaucoup plus de phrases qui ont des poids supérieurs ou égaux à 5 : 96 phrases, leurs longueurs se trouvent entre 17 et 83.

En regardant toutes ces 96 phrases, on constate que ces phrases contiennent souvent des *parataxis* (annotées en *dep* dans le treebank) et des *conj*, comme le cas du français et de l'anglais ; on a constaté également qu'il y a beaucoup d'erreurs d'annotation. Notamment, la relation *dep* et la relation *acl*[52] ne correspondent pas aux critères du guide d'annotation UD.

La figure 75 montre une de ces phrases dont l'annotation n'a pas été construite de façon bien conforme au guide. Notamment, nous trouvons que les relations *acl* sont annotées entre 都是 (*doushi* VERB) et 長大 (*zhangda* VERB), et entre 授意 (*shouyi* VERB) et 下 (*xia* ADP) ; la *parataxis* est remplacée par *dep* : entre 傳聞 (*chanwen* VERB) et 否認 (*foren* VERB) ; de plus, la direction de la *parataxis* est mal choisie : en effet selon le guide d'annotation, la relation *parataxis* devrait être à partir de l'élément à gauche vers l'élément à droite.

La figure 76 montre cette phrase corrigée. Pour cette version, les poids du flux diminuent considérablement. Pour montrer la différence, nous avons marqué la position (五 *wu*, 穀 *gu*) dans ces deux arbres. Le poids du flux est changé de 5 à 3 :

La position (五 *wu*, 穀 *gu*) dans l'arbre de la figure 75 :

- Niveau poids 1 : {五 <-*nummod*- 穀}
- Niveau poids 2 : {吃 -*obl*-> 雜糧, 吃 <-*acl*- 長大}
- Niveau poids 3 : {都是 -*xcomp*-> 長大, 都是 <-*acl*- 生病}
- Niveau poids 4 : {人 <-*nsubj*- 生病}
- Niveau poids 5 : {否認 -*ccomp*-> 不是}

[52] "*acl stands for finite and non-finite clauses that modify a nominal. The acl relation contrasts with the advcl relation, which is used for adverbial clauses that modify a predicate. The head of the acl relation is the noun that is modified, and the dependent is the head of the clause that modifies the noun.*" (*https://universaldependencies.org/u/dep/acl.html*)

Pour l’arbre de la figure 76, la même position :

- Niveau poids 1 : {五<-*nummod*- 穀}
- Niveau poids 2 : {吃 -*obj*-> 雜糧, 吃 <-*advcl*- 長大}
- Niveau poids 3 : {人 <-*nsubj*- 生病, 否認 -*parataxis*-> 生病}

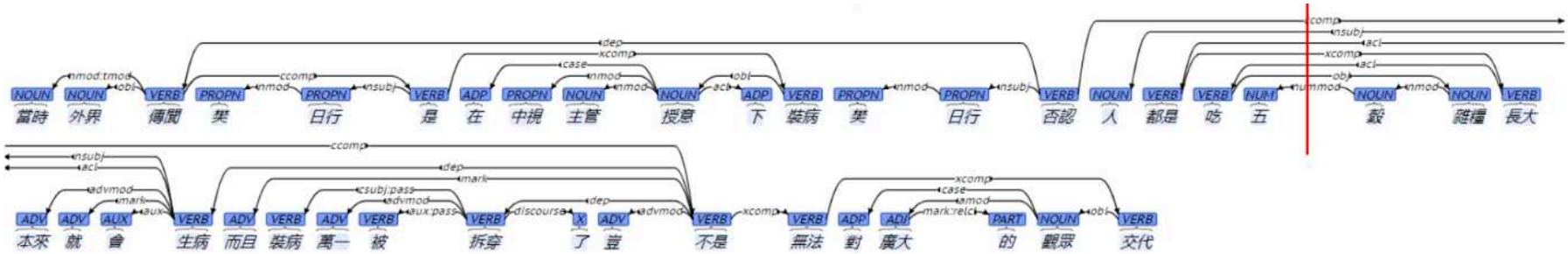

Figure 75 Exemple (UD_Chinese-(GSD), train-s10)

“當時外界傳聞樊日行是在中視主管授意下裝病,樊日行否認:「人都是吃五穀雜糧長大,本來就會生病;而且裝病萬一被拆穿了,豈不是無法對廣大的觀眾交代?”

當時 外界 傳聞 樊 日行 是 在 中視 主管 授意 下 裝病

Dang shi wai jie chuan wen Fan Ri xing shi zai zhong shi zhu guan shou yi xia Zhuang bing

À ce moment-là extérieur rumeur Fan Ri xing être à CTV directeur demander PART faire semblant d’être malade

À ce moment-là, la rumeur disait que Fan Rixing faisait semblant d'être malade à la demande du directeur de CTV.

樊　日行　否認：

Fan　Ri xing　fou ren

Fan　Ri xing　nier

Fan Rixing a nié.

人　都是　吃　五　穀　雜糧　長大，本來　就　會　生病

Ren　dou shi　chi　wu　gu　za liang　zhang da　ben lai　jiu　hui　sheng bing

Gens　tous être　manger　cinq　grains　céréale　grandir　ainsi　alors　pourra　être malade

Nous grandissons tous avec un régime alimentaire à base de céréales et de grains, il nous arrive de tomber malade.

(Il arrive à tout le monde de tomber malade.)

* wu　gu　zaliang :　*multi word expression*, signifie toutes les nourritures principales/basiques.

而且　　装病　　萬一　　被　　拆穿　　了，

Er qie　　zhuang bing　　wan yi　　bei　　chai chuan　le

De plus　faire semblant d’être malade　si jamais　être　démasquer　PART

豈　不是　　無法　　對　　廣大　　的　　觀眾　　交代　？

qi　bu shi　　zu fa　　dui　　guang da　　de　　guan zhong　　jiao dai

n’est-ce pas　impossible　à　large/grand/nombreux　PART　téléspectateurs　expliquer

Et si quelqu’un révèle qu’on a fait semblant d'être malade, on sera incapable de s’expliquer devant les nombreux téléspectateurs, n’est-ce pas ?

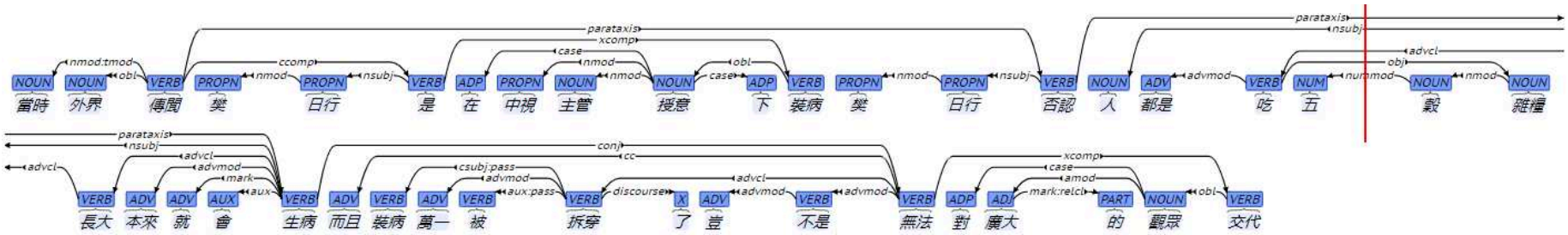

Figure 76. Version corrigée

5.3.2 Poids dans un corpus parlé vs. écrit

Dans le chapitre 2, section 2.4.3, nous avons discuté du corpus écrit et du corpus oral. Un travail a été réalisé par Yan (2017) pour examiner si la distribution des poids du flux diffère dans les corpus parlés et écrits. Nous avons considéré quatre treebanks en français : *UD_French*, *UD_French-ParTUT* et *UD_French-Séquoia* sont des treebanks en français écrit, et *UD_ French-Spoken* en français parlé. La figure 77 et le tableau 21 montrent la distribution des poids du flux dans ces treebanks.

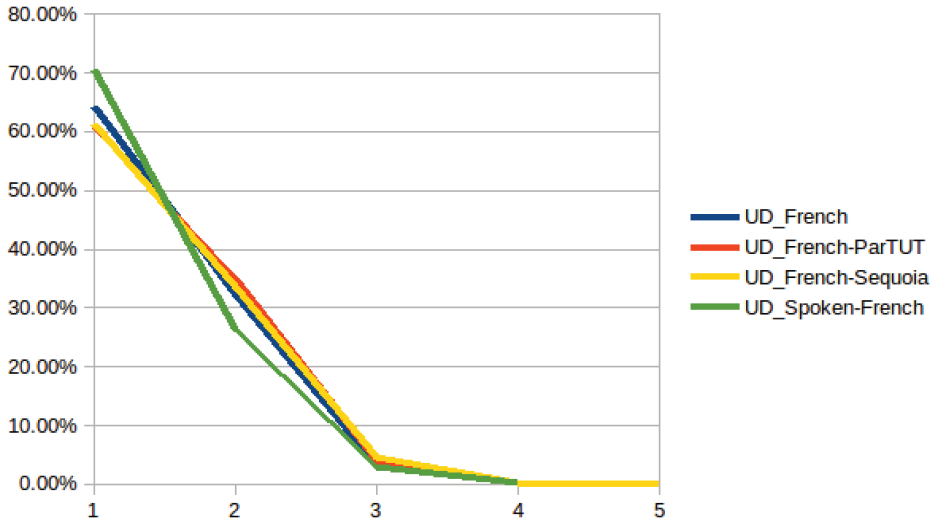

Figure 77. Distribution des poids du flux pour 4 treebanks en français (UD 2.0) (Yan, 2017)

	Tokens	Arbres	1	2	3	4	5
UD_French	392230	16031	64.31%	32.23%	3.27%	0.17%	0.01%
UD_French-ParTUT	17927	620	60.66%	34.99%	4.05%	0.26%	0.03%
UD_French-Sequoia	60574	2643	61.25%	33.76%	4.69%	0.29%	0.01%
UD_Spoken-French	33551	2636	70.54%	26.35%	2.92%	0.19%	0.00%

Tableau 21. Distribution des poids du flux pour 4 treebanks en français (UD 2.0) (Yan, 2017)

Par rapport aux corpus écrits, les résultats montrent que les corpus parlés ont des poids un peu plus faibles sauf pour les poids de 1. Plus précisément, 70% des positions de flux dans le corpus parlé *UD_French-Spoken* ont un poids égal à 1, contre 60,66% (*UD_French-ParTUT*), 64,31% (*UD_French(GSD)*) et 61,25% (*UD_French-Sequoia*) pour les corpus écrits. Le pourcentage de positions dans le corpus parlé ayant un poids

égal ou supérieur à 2 est inférieur à celui du corpus écrit. Et un poids de 5 est presque inexistant dans le corpus parlé. Le poids est donc plus contraint dans le corpus parlé que dans les trois corpus écrits. Par ailleurs on observe que les pourcentages sont robustes sur les trois corpus écrits où les courbes de la figure 77 sont pratiquement confondues.

5.3.3 Poids dans les arbres agrégés

Dans la section 2.4.2 du chapitre 2, nous avons discuté de la granularité (Kahane & Gerdes, 2021) de l'analyse pour les arbres syntaxiques. Les valeurs des poids, qui se trouvent de fait représenter les niveaux d'auto-enchâssement, varient en fonction de la granularité de l'analyse, puisque les unités minimales à prendre en compte sont différentes.

Les unités minimales utilisées dans les analyses de Lewis (1996) sont des sujets, des prédicats et des objets. Les résultats psycholinguistiques de Lewis (1996) montrent que le niveau d'auto-enchâssement pour une phrase est au maximum de 3 en anglais et de 4 en japonais. Dans le projet UD, les unités minimales pour les analyses syntaxiques sont des tokens. Par conséquent, le poids de flux maximum représentant le niveau d'auto-enchâssement trouvé par les expériences quantitatives ci-dessus est plus grand, en l'occurrence, on trouve en général 5. Cependant, nous savons que certaines dépendances dans l'arbre d'UD sont de nature fonctionnelle, et ses dépendants sont des mots grammaticaux. C'est le cas pour les dépendances comme *case* ou *det*. Les dépendances fonctionnelles sont également prises en compte dans le calcul du poids, mais ne sont pas nécessairement aussi complexes que les dépendances entre les mots pleins dans le traitement cognitif.

Dans les travaux de Yan et Kahane (2018), on a modifié la granularité des arbres syntaxiques avec un traitement d'agrégation. En conservant toutes les relations entre les mots pleins et en supprimant toutes les relations fonctionnelles (comme *case*, *det* et

aux etc.), cela revient à agréger les mots fonctionnels aux mots lexicaux. Nous obtenons une nouvelle version d'arbre pour la phrase que nous appelons l'arbre agrégé (angl : *aggregated tree*). Dans ce cas, les unités minimales sont des groupes des mots (voir aussi la partie ***Deux types de données*** pour les détails). L'hypothèse avancée dans les paragraphes précédents était bien qu'en tenant compte des unités minimales appropriées, les valeurs du poids peuvent être plus proche entre langues.

On a utilisé 37 treebanks en différentes langues. Ces treebanks sont disponibles sur le site de la conférence *Measuring linguistic complexity 2018*, ils sont sélectionnés dans la version 2.1 du projet UD. (Voir le détail : http://www.christianbentz.de/MLC_proceedings.html)

Deux types de données :

La figure 78 et la figure 79 montre ces deux types d'arbre pour la même phrase. L'un est l'arbre original et l'autre est l'arbre agrégé ayant une granularité plus grossière. Les unités minimales de l'arbre agrégé sont des groupes des mots, contrairement à l'arbre original, dont les unités minimales sont des *tokens* : par exemple, la dépendance *at* <– ***case-*** *Newsday* dans l'arbre original disparaît dans l'arbre agrégé, car l'unité minimale est devenue plus grande, c'est le groupe : *at Newsday*.

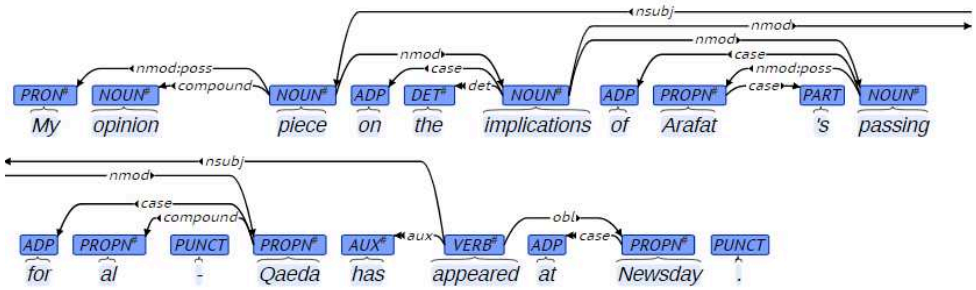

Figure 78 Arbre original (UD_English-EWT, weblog-juancole.com_juancole_20041120060600_ENG_20041120_060600-0001)

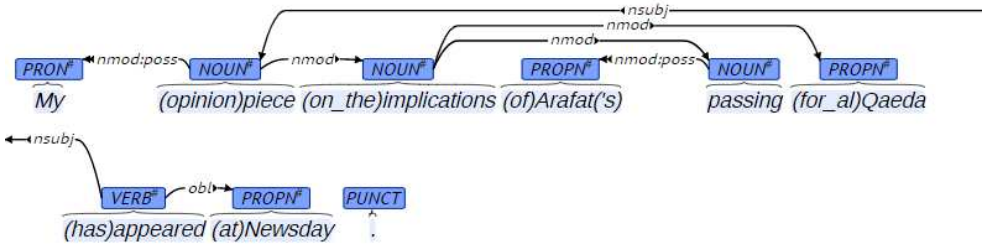

Figure 79 Arbre agrégé

Résultats dans 37 treebanks

En calculant le poids de toutes les positions du flux pour chaque treebank (voir les résultats dans le tableau 22), nous avons constaté que le poids maximum varie entre 3 et 5 dans les arbres agrégés. Précisément, il est de 3 pour le vietnamien et le slovaque, 4 pour 25 langues et 5 pour 10 langues. Par rapport aux treebanks originaux, où le poids maximum varie entre 4 et 6 : 4 langues ont des poids maximums égaux à 4, 21 langues pour 5 et 12 langues pour 6.

Les résultats pour le poids moyen sont similaires. Nous pouvons remarquer que le poids moyen est entre 1,67 (le polonais) et 1,18 (l'arabe) dans les arbres originaux contre 1,48 (l'urdu) et 1,10 (le polonais) dans les arbres agrégés.

Ainsi, les arbres agrégés ont des poids en borne supérieure moins dispersés (poids à 4 pour 25 langues) et plus faibles sur l'ensemble des langues des corpus.

Langue	Arbres originaux		Arbres agrégés	
	Poids max	Poids moyenne	Poids max	Poids moyenne
Afrikaans	5	1,53	4	1,33
Arabe	5	1,67	5	1,40
Basque	5	1,36	4	1,26
Bulgare	5	1,28	4	1,15
Catalan	6	1,48	4	1,25
Chinois	6	1,65	5	1,41
Croate	5	1,39	4	1,20
Tchèque	6	1,37	5	1,23
Danois	4	1,34	4	1,20
Néerlandais	5	1,40	5	1,22
Anglais	5	1,35	4	1,16
Estonien	6	1,29	4	1,17
Finnois	6	1,31	5	1,20
Français	5	1,39	4	1,19
Galicien	5	1,41	4	1,27
Grec	5	1,42	4	1,25
Hébreu	5	1,47	4	1,20
Hindi	6	1,58	5	1,42
Hongrois	6	1,54	5	1,33
Italien	5	1,39	5	1,21
Letton	6	1,35	5	1,19
Norvégien-Bokmaal	5	1,30	4	1,15
Norvégien-Nynorsk	6	1,32	4	1,16
Persan	6	1,64	4	1,46
Polonais	4	1,18	4	1,10

Portugais	5	1,43	4	1,22
Roumain	6	1,40	4	1,22
Russe	5	1,36	4	1,18
Serbe	5	1,40	4	1,21
Slovaque	4	1,18	3	1,09
Slovène	5	1,30	4	1,17
Espagnol	5	1,48	4	1,25
Suédois	5	1,32	4	1,17
Turc	6	1,48	4	1,34
Ukrainien	5	1,38	4	1,19
Urdu	5	1,66	5	1,48
Vietnamien	4	1,26	3	1,13

Tableau 22 Distribution des poids dans 37 treebanks (arbres originaux et arbres agrégés)

5.3.4 Commentaires

Les études quantitatives de cette section ont d'abord montré que le poids du flux est borné, et que la valeur maximale des poids du flux est entre 4 et 6 dans les treebanks UD. Ce qui n'est pas le cas pour la taille du flux, qui n'est pas bornée maximalement. Par ailleurs la taille montre une tendance à la minimisation dans les corpus étudiés. Le poids représente le niveau d'auto-enchâssement des constructions et ne peut pas augmenter sans limite comme nous l'avons déjà commenté. Cela est probablement dû à la limitation de la mémoire de travail.

Ensuite, le poids peut varier dans le corpus écrit et le corpus oral. Les résultats des treebanks en français ont montré que le poids dans le corpus parlé est plus contraint que dans le corpus écrit. Nous avons observé au niveau qualitatif dans les exemples analysés précédemment que de nombreuses phrases dans les corpus écrits ont des longues

dépendances *conj* ou *parataxis*, ce qui augmente le poids. Par ailleurs, on pense que cette différence du poids pourrait également être due à la nature non planifiée et spontanée de la langue parlée. La non-planification et la spontanéité nécessitent une plus grande consommation de mémoire de travail, ce qui conduit en général à des constructions syntaxiques moins complexes.

Enfin, la discussion sur le poids dans les arbres agrégés se divise en deux parties :

- Tout d'abord, les poids maximaux dans les arbres agrégés des différentes langues sont considérablement semblables. Dans les arbres agrégés, les dépendances fonctionnelles sont supprimées et seules les dépendances entre les mots pleins sont considérées. Ainsi, les arbres agrégés relèvent plus du niveau sémantique que du niveau syntaxique. De plus, nous savons que les mots grammaticaux sont utilisés différemment dans des langues, et même que certaines dépendances fonctionnelles ne sont pas présentes dans toutes les langues. Cela accentue également les différences de poids de flux entre les langues. Par conséquent, les arbres agrégés permettent d'obtenir naturellement des résultats plus homogènes.
- Nous allons mettre maintenant l'accent sur le lien entre les résultats et les études psychologiques. Nous avons expliqué au chapitre 1 que la limitation de la mémoire de travail est d'environ 4 *chunks* (voir la définition du *chunk* de Cowan (2004) et dans la section 1.2.3 du chapitre 1). Le poids du flux représente le nombre maximum des dépendances disjointes. Et les dépendances disjointes ne peuvent pas être traitées par la mise en facteur de l'information comme les dépendances en bouquet (voir la section 4.2.2 du chapitre 4). Elles seraient stockées dans la mémoire de travail indépendamment. Le poids du flux dans les arbres agrégés est également autour de 4. Nous faisons l'hypothèse que la limitation du poids du flux soit effectivement liée à la limitation de la mémoire de travail.

5.4 Flux potentiel projectif

Cette section se concentre sur des études quantitatives du flux potentiel (projectif). Nous avons déjà discuté du flux potentiel dans la section 4.3.4 du chapitre 4. Nous rappelons ici que le flux potentiel dans une position inter-mot donnée est l'ensemble des mots avant cette position qui sont susceptibles d'être liés à un mot après celle-ci tout en respectant la projectivité (Kahane et Yan, 2019).

Nous allons présenter les études quantitatives du flux potentiel projectif en deux parties.

La section 5.4.1 pose la question sur la relation entre le flux potentiel et la complexité syntaxique. Plus précisément, en observant la distribution des valeurs du flux potentiel dans des treebanks, nous pouvons fouiller pour voir si le flux potentiel est contraint. L'hypothèse de départ pour établir le flux potentiel projectif est que la structure de la phrase est projective. Comme l'interlocuteur traite la phrase de gauche à droite tout en respectant la projectivité, les informations qui peuvent être utilisées par la suite seront stockées dans la mémoire de travail. Par contre, sans la contrainte de projectivité, on ne peut pas prévoir la structure qui suivra, car tous les éléments qui ont été traités peuvent établir des dépendances avec la suite. Cela augmente considérablement le contenu du stockage de travail. Nous pensons donc que la taille du flux potentiel ne peut être extrêmement importante, car la capacité de stockage dans la mémoire de travail est limitée. De ce fait, nous allons calculer et analyser les tailles du flux potentiel dans SUD. Ensuite, nous allons également comparer la distribution des tailles du flux potentiels avec celle des tailles du flux observé. Le flux potentiel pour une position enregistre le nombre des mots à gauche de la position à partir de la construction partielle, alors que le flux observé pour une position capture l’ensemble des dépendances à travers la position à partir de l’arbre syntaxique complet. Il est important de noter que la taille du flux observé n'est pas forcément plus petite que celle du flux potentiel à une position donnée, car un mot dans le flux potentiel peut établir des dépendances avec plusieurs mots à droite.

Dans la section 5.4.2, nous examinerons précisément la distribution du flux potentiel en fonction du type de langue. Comme expliqué dans la section 4.3.4 du chapitre 4, la distribution des valeurs du flux potentiel dans les treebanks serait différenciée selon le type de langue (voir également 2.4.4 sur le type de langue).

5.4.1 Flux potentiel dans les treebanks SUD

Le tableau 23 montre la distribution des valeurs du flux potentiel en détail, et nous permet d'analyser sa valeur contrainte. Nous avons constaté qu'il est nécessaire d'atteindre la taille 20 pour avoir plus de 99,9% des positions du flux potentiel, alors que cette valeur est à 10 pour le flux observé (voir 8.2.1 pour les résultats en détail du flux observé).

n	% flux potentiel projectif= n	% flux potentiel projectif ≤n
1	15,3372	15,3372
2	19,2018	34,5390
3	17,5429	52,0819
4	13,8746	65,9565
5	10,3640	76,3205
6	7,4554	83,7759
7	5,2398	89,0158
8	3,6044	92,6202
9	2,4420	95,0622
10	1,6400	96,7022
11	1,0883	97,7905
12	0,7246	98,5151
13	0,4769	98,9920
14	0,3182	99,3102

15	0,2124	99,5226
16	0,1434	99,6660
17	0,0972	99,7633
18	0,0663	99,8296
19	0,0460	99,8756
20	0,0323	99,9079

Tableau 23. Distribution des tailles du flux potentiel projectif dans SUD 2.7

La figure 80 et la figure 81 comparent la distribution des tailles du flux potentiel avec celle des tailles du flux observé pour tous les treebanks SUD. La distribution des valeurs du flux potentiel est plus plate : 19,20 % contre 32,08 % des positions pour une taille du flux observé égale à 2 ; à partir de la taille 2, la fréquence du flux potentiel diminue, mais plus lentement que celle pour le flux observé. Cela signifie que le flux potentiel projectif a généralement des tailles plus importantes que le flux observé.

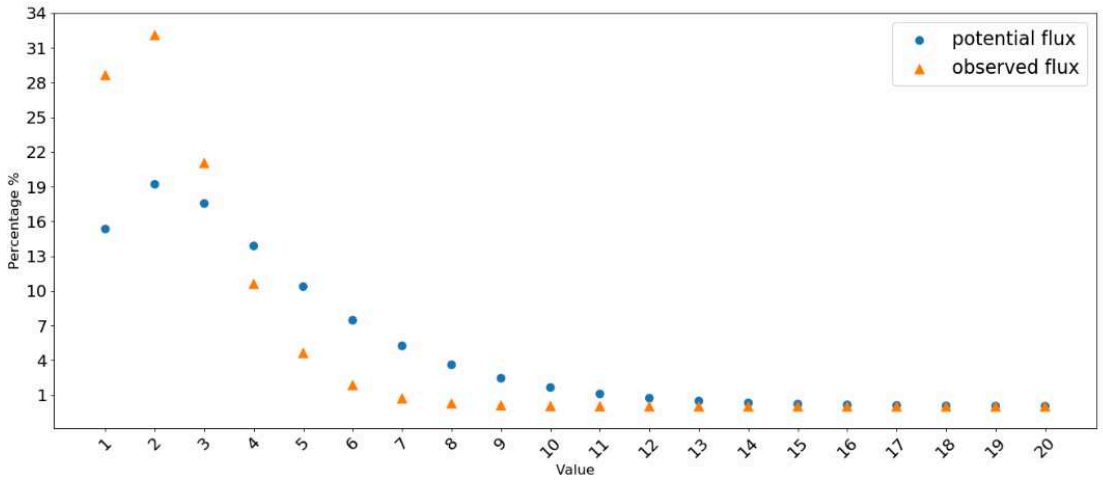

Figure 80 Répartition des tailles du flux potentiel (projectif) et des tailles du flux observé dans SUD 2.7

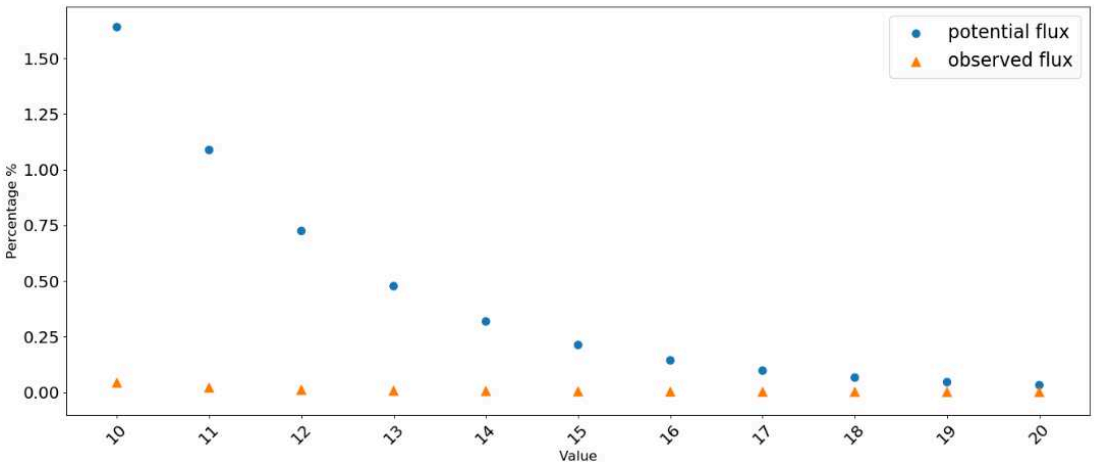

Figure 81 Répartition des valeurs entre 10 et 20 pour la taille du flux potentiel (projectif) et la taille du flux observé dans SUD 2.7

5.4.2 Flux potentiel dans les treebanks de différentes langues

Dans 4.3.4, nous avons expliqué que le calcul du flux potentiel (projectif) n'est pas le même selon la direction de la dépendance. Si le gouverneur se trouve devant son dépendant, ils sont tous deux accessibles pour établir de nouvelles dépendances avec les mots qui suivent, tout en respectant la projectivité. Les deux éléments font partie du flux potentiel. Dans le cas du gouverneur après son dépendant, seul le gouverneur est accessible en maintenant la projectivité.

Dans les langues à tête initiale, les dépendants se placent après leur gouverneur et inversement dans les langues à tête finale. Ainsi, nous nous attendons à ce que le flux potentiel soit assez différent selon que la langue est à tête initiale ou à tête finale.

Nous reprenons les résultats typométriques de Gerdes, Kahane et Chen (2021) dans la figure 22 de la section 2.4.4. Nous considérons les intervalles 0 à 20 % et 80 à 100 % pour déterminer les langues à tête finale et les langues à tête initiale. Cela comprend respectivement 12 puis 2 langues：

- Langues à tête finale : japonais, coréen, télougou, tamoul, kazakh, ouïgour, turc, marathi, bouriate, hindi, ourdou et sanscrit.

- Langues à tête initiale : arabe et irlandais.

La figure 82 montre les flux potentiels dans les treebanks SUD des deux types de langues. Nous trouvons le tableau 24 pour présenter précisément les données. Celui-ci comprend la distribution des valeurs égales à *n* pour les positions du flux potentiel, ainsi que la distribution pour les valeurs égales et inférieures à *n*.

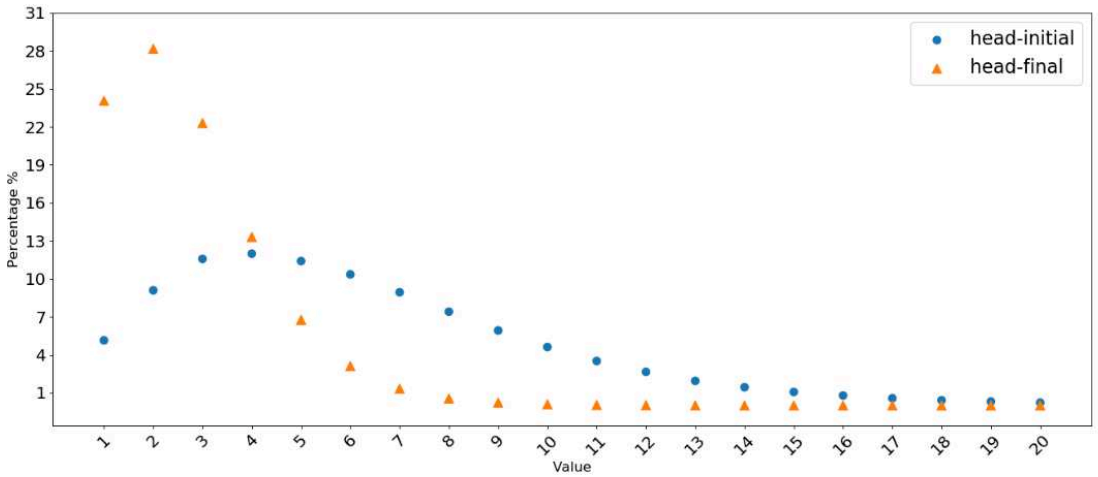

Figure 82. Distribution des tailles du flux potentiel projectif pour l'ensemble des langues à tête finale et celui des langues à tête initiale (SUD 2.7)

n	% flux potentiel projectif= n		% flux potentiel projectif ≤n	
	Langues initiales	Langues finales	Langues initiales	Langues finales
1	5,1397	24,0602	5,1397	24,0602
2	9,0932	28,1527	14,2329	52,2129
3	11,5746	22,2911	25,8076	74,5040
4	11,9855	13,3051	37,7930	87,8092
5	11,3998	6,7633	49,1928	94,5724
6	10,3471	3,1130	59,5399	97,6855
7	8,9344	1,3277	68,4743	99,0132

8	7,4074	0,5544	75,8817	99,5676
9	5,9247	0,2278	81,8064	99,7954
10	4,6184	0,0995	86,4248	99,8950
11	3,5150	0,0458	89,9398	99,9408
12	2,6528	0,0237	92,5925	99,9645
13	1,9400	0,0118	94,5325	99,9762
14	1,4446	0,0071	95,9771	99,9833
15	1,0717	0,0045	97,0488	99,9878
16	0,7902	0,0032	97,8390	99,9909
17	0,5721	0,0027	98,4112	99,9936
18	0,4191	0,0014	98,8303	99,9950
19	0,3120	0,0010	99,1423	99,9960
20	0,2268	0,0007	99,3691	99,9968
21	0,1627	0,0006	99,5318	99,9973
22	0,1180	0,0005	99,6498	99,9978
23	0,0905	0,0004	99,7403	99,9982
24	0,0674	0,0002	99,8077	99,9985
25	0,0495	0,0003	99,8572	99,9987
26	0,0345	0,0003	99,8916	99,9990
27	0,0259	0,0003	99,9175	99,9993

Tableau 24. Distribution des tailles du flux potentiel projectif pour l'ensemble des langues à tête finale et des langues à tête initiale (SUD 2.7)

Nous voyons bien que les distributions diffèrent entre les deux types de langues. Pour les langues à tête finale, la distribution des tailles du flux potentiel est similaire à celle pour l'ensemble des treebanks SUD (Figure 80). Mais ce n'est pas le cas pour les langues à tête initiale. Nous pouvons constater que pour les langues à tête finale, les valeurs de flux potentiel de 1, 2, et 3 sont nettement plus importantes que celles pour les langues à tête initiale : 5,14%, 9,09% et 11,57% pour l'ensemble des langues à tête

initiale contre 24,06%, 28,15% et 22,29% pour l'ensemble des langues à tête finale.

Au total, 99,9% des positions du flux potentiel ont une taille inférieure à 27 pour l'ensemble des langues à tête initiale. Et cette valeur est de 11 pour l'ensemble des langues à tête finale.

Donc, les tailles du flux potentiel dans les langues à tête initiale sont plus grandes par rapport aux langues à tête finale.

Flux potentiel dans cinq langues choisies

Pour examiner encore plus en détail les distributions dans les langues, nous allons présenter les distributions des tailles du flux potentiel de deux langues à tête initiale (l'arabe et l'irlandais) et celles de trois langues à tête finale (le truc, l'hindi et le japonais). La figure 83 et figure 84 montrent la distribution du flux potentiel dans ces cinq langues.

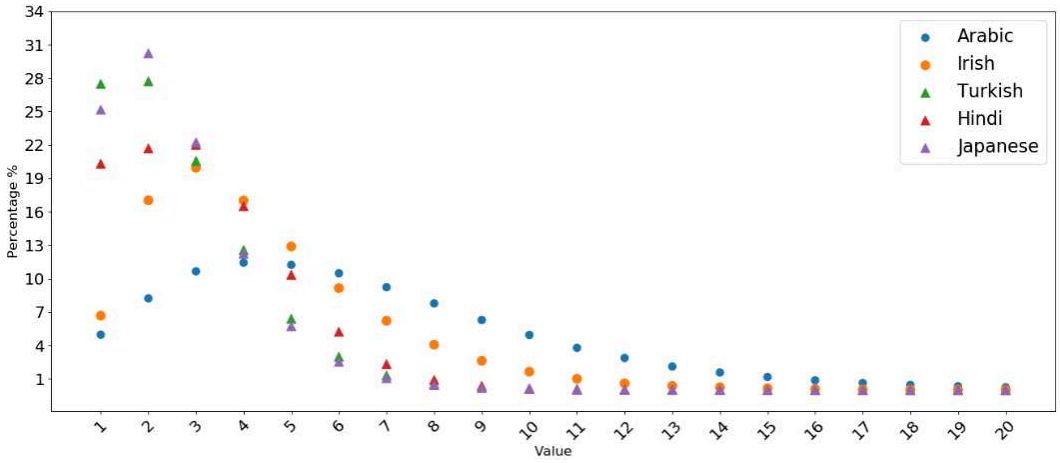

Figure 83 Distribution des tailles du flux potentiel dans 5 langues (SUD 2.7)

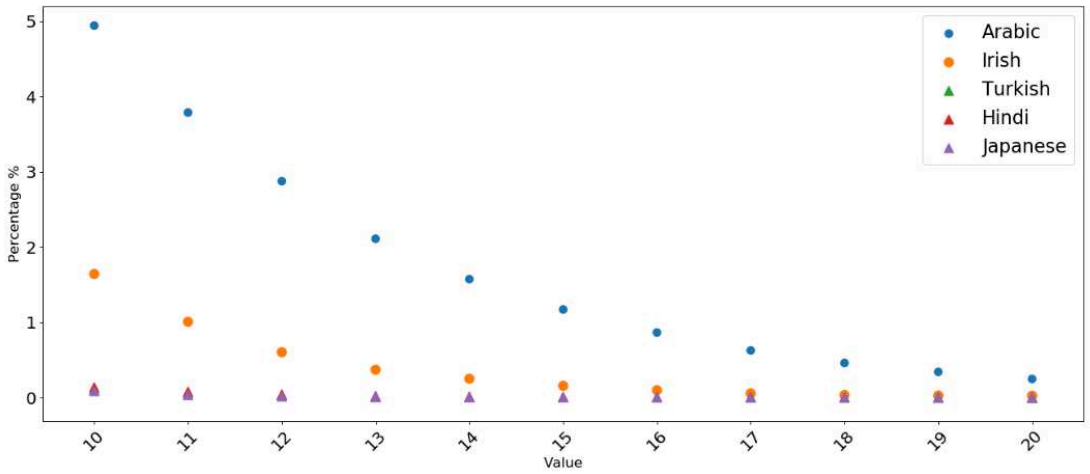

Figure 84 Distribution des tailles du flux potentiel dans 5 langues, valeurs entre10 et 20 (SUD 2.7)

Comme on l'a appris dans la partie précédente, on peut bien constater que les trois langues à tête finale ont des tailles plus petites par rapport aux deux langues à tête initiale (voir aussi 8.2.2 pour les résultats en détail).

Par ailleurs, la longueur de la phrase peut avoir un effet sur la distribution de la taille du flux potentiel. Plus la phrase est longue, plus il est probable que la taille du flux potentiel soit plus important. Nous constatons que des petites valeurs du flux potentiel pour le turc sont nettement plus nombreuses que pour toutes les autres langues. En effet, la longueur moyenne des phrases dans les treebanks du turc est seulement de 8,08 : c'est probablement parce que le turc est une langue très agglutinante avec beaucoup de morphèmes par mot et globalement pas de mots grammaticaux. Alors que cette longueur pour les autres langues, est de 20,4 pour le japonais, 19,8 pour l'hindi, 33,2 pour l'arabe et 21,04 pour l'irlandais.

Comparaison avec la taille du flux observé

La figure 85 et la figure 86 montrent la taille du flux pour ces cinq langues. Les résultats dans l'ensemble des données SUD présentés précédemment montrent que le flux potentiel a tendance à avoir des tailles plus importantes que le flux observé, cette tendance est notamment accentuée dans les langues à tête initiale.

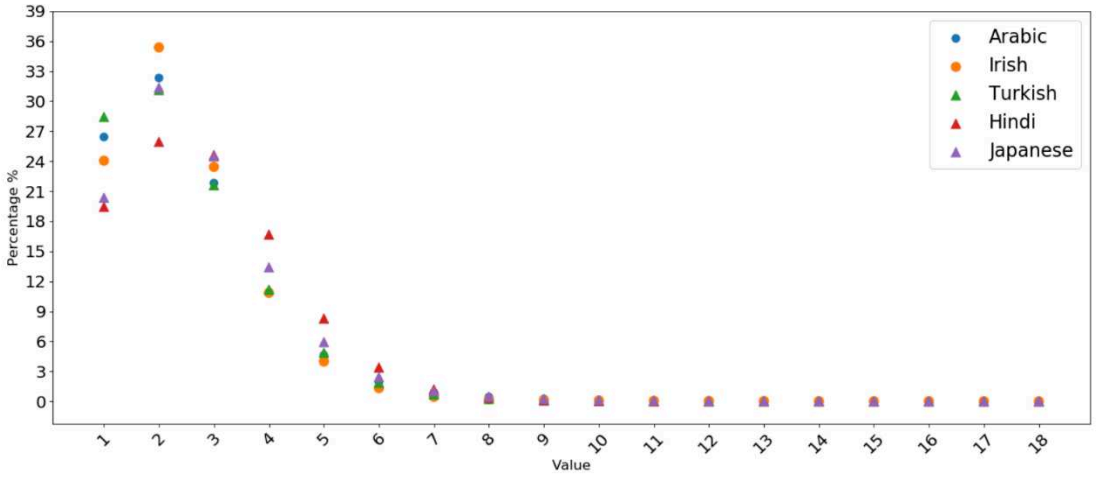

Figure 85. Distribution des tailles du flux observé dans 5 langues (SUD 2.7)

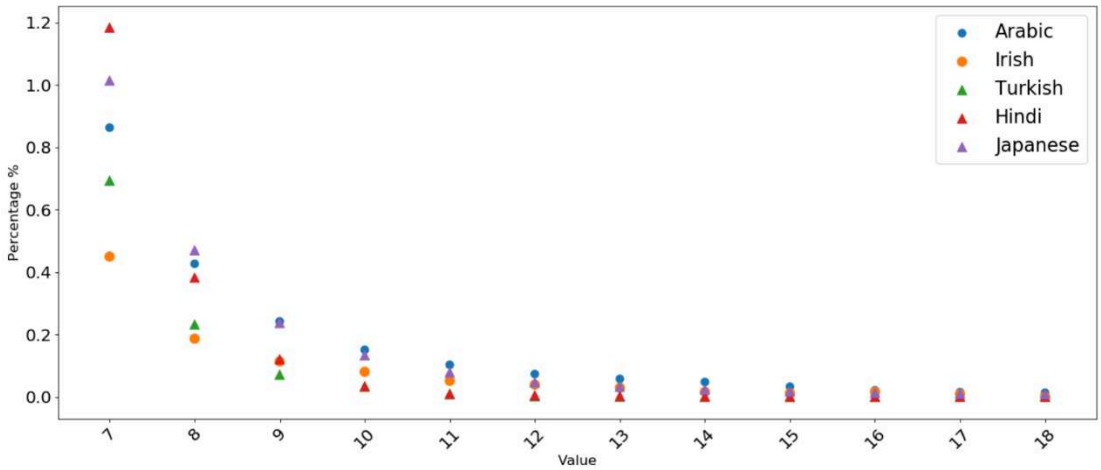

Figure 86. Distribution des tailles du flux observé dans 5 langues, valeurs entre10 et 18 (SUD 2.7)

Les positions du flux potentiel de petite taille sont beaucoup moins nombreuses que celles du flux observé (voir aussi 8.2.3 pour les données des résultats) :

- Pour l'arabe, 4,97 %, 8,22 % et 10,66 % des positions du flux potentiel de taille 1, 2 et 3 contre 26,42%, 32,31 % et 21,81 % pour le flux observé.
- Pour l'irlandais, 6,68 %, 17,03 % et 19,96 % des positions du flux potentiel de taille 1, 2 et 3 contre 24,05%, 35,37 % et 23,43 % pour le flux observé.
- 76,14 % des positions du flux potentiel de taille ≥ 4 contre 19,46% des positions du flux observé pour l'arabe, et 56,34% contre 17,15 % pour l'irlandais.

5.4.3 Commentaires

Les résultats du flux potentiel dans l'ensemble des treebanks SUD montrent que la taille du flux potentiel est contrainte. Nous considérons le flux potentiel comme une métrique qui peut également être utilisée pour évaluer la difficulté de traitement des phrases, compte tenu justement des éléments de contrainte.

Dans les treebanks UD en différentes langues, on a également examiné la relation entre le flux potentiel et le type de langue. Les résultats ont montré que les tailles du flux potentiel sont plus importantes dans les langues à tête initiale que dans les langues à tête finale. Ceci est conforme à notre hypothèse selon laquelle les différences dans la distribution de la taille du flux potentiel peuvent être déclinées par type de langue.

Nous avons comparé aussi la distribution du flux potentiel avec celle du flux observé dans les deux types des langues. Dans les langues à tête finale, la distribution du flux potentiel est similaire à celle du flux observé. Par contre, dans les langues à tête initiale le flux potentiel est plus important que dans les langues à tête finale et le flux observé lui est plus faible. Ces différences permettent d'expliquer pourquoi leurs structures dans les langues à tête initiale sont plus contraintes en termes du flux observé et donc ont une complexité moins importante :

- Pour les langues à tête initiale, il est beaucoup plus difficile de prévoir quel nœud du flux potentiel établira une dépendance avec le nœud suivant lors du traitement syntaxique de gauche à droite. Ainsi, leurs arbres syntaxiques ont tendance à avoir des contraintes sur leur structure, qui provoque moins de flux observé en raison de l'impact du flux potentiel plus important.
- Pour les langues à tête finale, au contraire, les tailles du flux potentiel sont relativement faibles. Ce dernier est donc plus prévisible pour les structures arborescentes. Par conséquent, les langues à tête finale peuvent avoir moins de contrainte sur leur structure, et donc des tailles de flux observées relativement importantes.

5.5 Autres expériences

Dans cette section, nous allons présenter trois autres expériences. Il s'agit tout d'abord d'une étude sur la distribution des différentes dépendances disjointes. L'objectif est d'examiner si la complexité syntaxique diffère selon le type de dépendance dans les dépendances disjointes. Ensuite, nous présenterons brièvement une étude sur le poids combiné qui est une variante du poids du flux. Finalement, nous présenterons des études sur le flux requis qui ont obtenu des conclusions similaires à celles pour le flux potentiel.

5.5.1 Dépendances disjointes

Dans cette section, nous présentons notre étude préliminaire sur les différentes configurations de dépendances disjointes. L'objectif est de découvrir s'il existe des situations particulières en fréquence pour les ensembles de dépendances disjointes concomitantes. Comme ce que nous avons évoqué dans la section 4.3.2.3 du chapitre 4, nous nous focalisons ici sur l'étude des dépendances disjointes contenant le sujet (*nsubj*) et sur celles contenant l'objet (*obj*).

Nous allons proposer quatre configurations parmi tous les groupes de dépendances disjointes pour notre étude, ce sont celles de la figure 87.

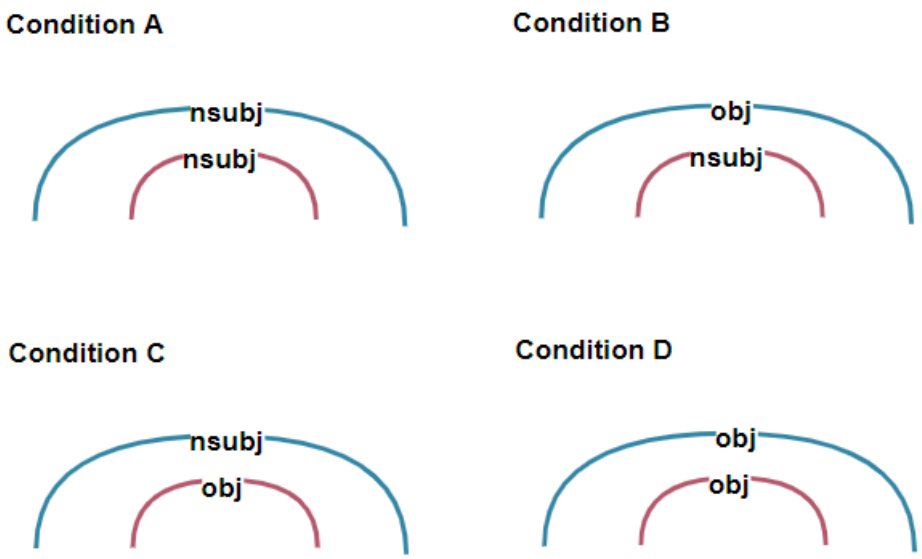

Figure 87 Quatres conditions pour deux dépendances disjointes (nous considérons ici seulement la dépendance du type objet (*obj*) et la dépendance du type sujet (*nsubj*))

Nous pouvons représenter ces quatre conditions sous forme matricielle comme dans le tableau 25, où les lignes représentent les dépendances en dessous et les colonnes représentent les dépendances au-dessus.

Bas / ***Haut***	***nsubj***	***obj***
nsubj	$S_{bas}S_{haut}$	$S_{bas}O_{haut}$
obj	$O_{bas}S_{haut}$	$O_{bas}O_{haut}$

Tableau 25 Représentation de quatre conditions (*S: nsubj, O: obj*)

Selon l'hypothèse sur l'idée d'**interférence** de Lewis et Vasishth (2005) (voir 4.3.2 et 1.4.3.1), nous pensons que deux dépendances disjointes similaires seront difficiles à traiter par rapport à deux dépendances disjointes différentes, car la similitude pourrait provoquer des interférences pendant le traitement cognitif. Pour les cas dans lesquels les deux dépendances disjointes sont toutes deux *nsubj* ou toutes deux *obj*, donc les deux situations de similarité parmi les quatre proposées, nous attendons une fréquence d'apparition faible. En effet, on pourrait avoir tendance à éviter d'utiliser de telles structures car elles sont plus difficiles à traiter.

Résultats

Comme nous l'avons expliqué à la section 4.3.2.3 du chapitre 4, nous allons mettre à disposition deux types de données : l'une concerne les nombres d'occurrences, calculées dans le corpus ; l'autre concerne les nombres attendus d'occurrences, prédits sur la base du rapport *a* entre la probabilité d'occurrences de *nsubj* et celle de *obj*. Cette prédiction présuppose que les occurrences de la dépendance en bas et de la dépendance en haut sont considérées comme des événements indépendants. Nous reprenons ici le tableau 26 pour rappeler les éléments nécessaires au calcul des valeurs attendues.

B / ***H***	***nsubj***	***obj***
nsubj	$O_{bas}O_{haut} \times a^2$	$O_{bas}O_{haut} \times a$
obj	$O_{bas}O_{haut} \times a$	$O_{bas}O_{haut}$

Tableau 26 Nombres d'occurrences attendus ($a = \frac{nombre\ d'occurences\ nsubj}{nombre\ d'occurence\ obj}$)

Résultats dans l'ensemble des données de UD

Le tableau 27 montre les nombres d'occurrences attendus à partir de la valeur a. Le tableau 28 présente les nombres d'occurrences expérimentales, avec entre parenthèses leur pourcentage de diminution ou d'augmentation par rapport au nombre attendu.

B / ***H***	***nsubj***	***Obj***
nsubj	57469	40591
obj	40591	28670

Tableau 27 Nombres d'occurrences attendus ($a = \frac{N_{nsubj}}{N_{obj}} = \frac{1759487}{1242743} = 1{,}415809222$)

B / ***H***	***nsubj***	***obj***
nsubj	46868 (↓18,45%)	**11141 (↓72,55%)**
obj	**54044 (↑33,14%)**	28670

Tableau 28 Nombres d'occurrences observés

Dans l'ensemble des données de UD, le *nsubj* est plus fréquent que l'*obj*. Il s'agit de 1759487 occurrences de *nsubj* contre 1242743 d'*obj.* Ainsi, nous obtenons une valeur pour a de 1,42.

On observe que $S_{bas}S_{haut}$ et $S_{bas}O_{haut}$ sont moins fréquents que prévu. Pour $S_{bas}S_{haut}$, il s'agit une diminution de 18,45% par rapport au nombre attendu ; pour $S_{bas}O_{haut}$, il s'agit d'une diminution de 72% par rapport au nombre attendu pour cette configuration.

Par ailleurs, le point plus important est que le contraste entre $S_{bas}O_{haut}$ et $O_{bas}S_{haut}$ est très fort. Cela montre que la probabilité de deux dépendances disjointes composées d'un *nsubj* et un *obj* dépend de leur emplacement en haut et en bas.

Pour mieux comprendre les raisons de ces remarques sur les résultats de l'ensemble de données UD, nous devons examiner les résultats des treebanks par langue.

Analyses dans les treebanks en français, anglais et chinois

Nous nous concentrons sur les résultats des langues qui nous sont aujourd'hui les plus proches c'est-à-dire, le français, l'anglais et le chinois. Le tableau 29, le tableau 30 et le tableau 31 montrent les résultats expérimentaux sur ces langues.

Bas \ ***Haut***	***nsubj***	***obj***
nsubj	1835 (↑744,52%)	**249 (↑97,71%)**
obj	**1461 (↑1060,05%)**	73

Tableau 29 Français – Nombres d'occurrences observés ($a = \frac{N_{nsubj}}{N_{obj}} = \frac{65116}{37743} = 1{,}725247066$)

Bas / ***Haut***	***nsubj***	***obj***
nsubj	1004 (↑827,62%)	**137 (↑125,84%)**
obj	**603 (↑894,02%)**	34

Tableau 30 Anglais – Nombres d'occurrences observés ($a = \frac{N_{nsubj}}{N_{obj}} = \frac{55692}{31214} = 1{,}784199398$)

Nous avons trouvé des résultats similaires pour le français et l'anglais. Tout d'abord, leurs valeurs pour a sont proches au centième (1,72 pour le français et 1,78 pour l'anglais). Deuxièmement, les trois conditions, $S_{bas}S_{haut}$, $S_{bas}O_{haut}$ et $O_{bas}S_{haut}$ sont plus fréquentes que prévu. Nous pensons que les dépendances disjointes avec S_{haut} sont plus probables que celles avec O_{haut}. Quant à l'occurrence de $O_{bas}O_{haut}$, c'est la plus rare dans les deux langues.

De plus, nous avons trouvé un phénomène similaire dans UD global : une très grande différence entre $S_{bas}O_{haut}$ et $O_{bas}S_{haut}$. Bien que les deux configurations soient plus élevées que prévu, $O_{bas}S_{haut}$ se produit beaucoup plus souvent que $S_{bas}O_{haut}$.

Bas / ***Haut***	***nsubj***	***obj***
nsubj	670 (↓17,30%)	344 (↓48,09%)
obj	2542 (↑283,62%)	542 (0%)

Tableau 31 Chinois- Nombres d'occurrences observés ($a = \frac{N_{nsubj}}{N_{obj}} = \frac{13018}{10648} = 1{,}22257701$)

Les résultats pour le chinois diffèrent de ceux des deux langues ci-dessus. Tout d'abord, nous constatons que le rapport *a* entre les occurrences de *nsubj* et d'*obj* est plus proche de 1, ce qui implique que la distribution de ces deux relations est plus équilibrée par rapport à celles pour le français et pour l'anglais.

Ensuite, nous constatons que les dépendances disjointes qui contiennent O_{bas} sont plus fréquentes que les autres cas. Pour toutes les dépendances disjointes qui contiennent

S_{bas}, S_{haut} apparaît plus fréquemment que O_{haut}. Le contraste entre $S_{bas}O_{haut}$ et $O_{bas}S_{haut}$ est toujours là : $S_{bas}O_{haut}$ a diminué de 48% par rapport au nombre attendu, tandis que $O_{bas}S_{haut}$ a augmenté de 283,62% par rapport au nombre attendu.

À partir des observations ci-dessus, nous concluons dans un premier temps que nos résultats varient en fonction de la langue :

- Dans les résultats pour les trois langues ci-dessus, pour l'anglais et le français, il est nettement plus difficile de placer un *obj* en haut par rapport au *nsubj*.
- En ce qui concerne le chinois, la situation n'est pas aussi claire. *obj* se trouve plus fréquent en bas par rapport au *nsubj.*
- Par ailleurs, deux relations identiques en disjointes ($S_{bas}S_{haut}$ et $O_{bas}O_{haut}$) ne sont toujours pas aussi rares que ce que l'on attendait.

Nous faisons l'hypothèse que les différentes distributions de ces quatre conditions sont dues à l'ordre des mots de la langue. Nous allons prendre des exemples dans les treebanks de ces trois langues pour discuter.

Les exemples dans les treebanks français, anglais et chinois[53]

Nous avons remarqué que la position de la proposition subordonnée dans ces trois langues est différente. En français et en anglais, la proposition subordonnée s'attache toujours à droite du nœud dont elle dépend. Comme le montre la figure 88 , lorsqu'une proposition subordonnée s'attache au sujet principal, une dépendance *nsubj* et/ou *obj* de la subordonnée est couverte par une dépendance *nsubj* entre le verbe principal et le sujet principal. Cela entraîne les dépendances disjointes en $S_{bas}\boldsymbol{S_{haut}}$ et/ou en $O_{bas}\boldsymbol{S_{haut}}$.

Cependant, lorsqu'une proposition subordonnée s'attache à l'objet principal, une

[53] Voir toutes les phrases extraites pour ces trois langues à : https://github.com/chunxiaoyan/results_experiments_phd

dépendance *obj* entre le verbe principal et l'objet principal ne peut couvrir aucun *nsubj* ou *obj* de la proposition subordonnée. De ce fait, on ne peut constituer des dépendances disjointes.

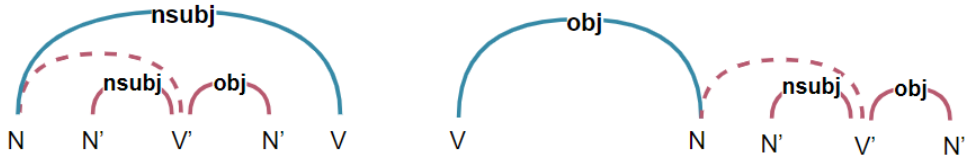

Figure 88 Propositions subordonnées en français et en anglais

Il en résulte que $S_{bas}\boldsymbol{S_{haut}}$ et $O_{bas}\boldsymbol{S_{haut}}$ sont nettement plus nombreuses que $S_{bas}\boldsymbol{O_{haut}}$ et $O_{bas}\boldsymbol{O_{haut}}$ dans les treebanks. Dans l'exemple de la figure 89, il s'agit d'une proposition relative qui s'attache au sujet principal *la question*. La subordonnée *que vous posez* se trouve au-dessous de la dépendance *nsubj*. Ainsi, on trouve deux configurations de dépendances disjointes en $S_{bas}\boldsymbol{S_{haut}}$ *et* $O_{bas}\boldsymbol{S_{haut}}$. En ce qui concerne l'exemple en anglais de la figure 90, la subordonnée *Reco can levy against an agent* se trouve entre le sujet principal et le verbe principal, ce qui forme deux dépendances disjointes en $S_{bas}\boldsymbol{S_{haut}}$.

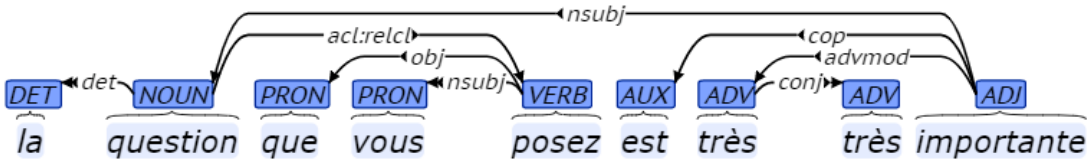

Figure 89 Exemple (UD_French-spoken, Rhap_D2009-12)

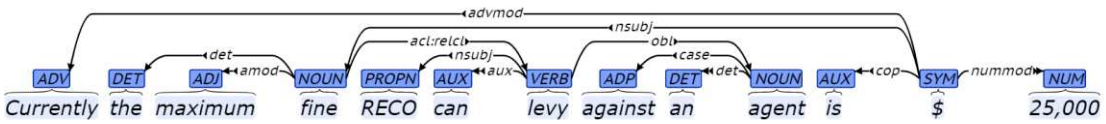

Figure 90 Exemple (UD_English-PUD, n01036033)

Pour les dépendances disjointes qui contiennent deux objets, nous n'avons trouvé que peu de cas. Nous montrons deux exemples ici, la figure 91 et la figure 92. Dans les exemples, on peut trouver qu'il ne s'agit pas d'une proposition subordonnée relative,

mais d'une proposition subordonnée adverbiale. Elle s'attache au verbe principal, et se trouve à gauche de l'objet principal. Ainsi cela crée deux dépendances en $O_{bas}O_{haut}$.

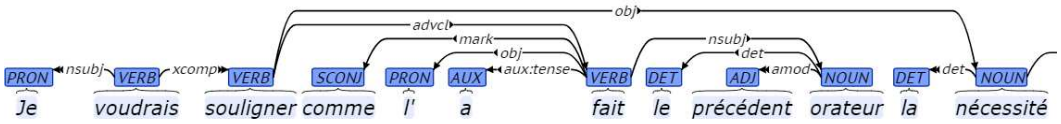

Figure 91 Exemple (UD_French-Squoia, Europar.550_00522)

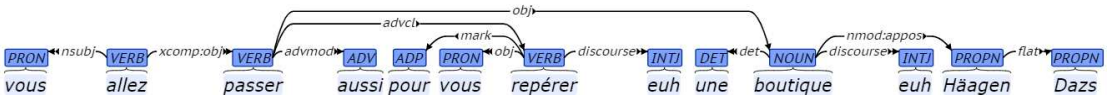

Figure 92 Exemple (UD_French-Spoken, Rhap_D0008-34)

En ce qui concerne le chinois, la proposition qui s'attache au sujet se trouve à gauche. Comme le montre la figure 93, il en résulte que la dépendance *nsubj* entre le verbe principal et le sujet principal ne couvre pas la proposition subordonnée qui s'attache au sujet principal. Cependant, celle d'*obj* couvre la proposition subordonnée qui s'attache à l'objet principale. Cela entraîne deux dépendances disjointes en $O_{bas}O_{haut}$ ou[54] $S_{bas}O_{haut}$.

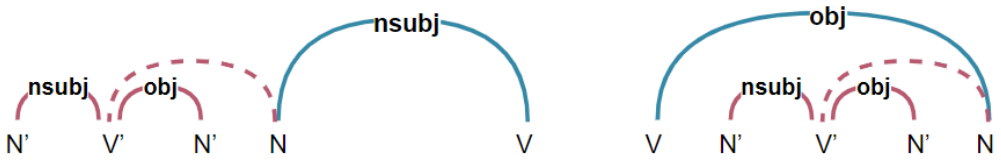

Figure 93 Propositions subordonnées en chinois

[54] En chinois, il n'y a pas de pronom relatif *qui* ou *que* comme en français. Dans la proposition relative avec l'extraction du sujet, il n'y a pas de dépendance *nsubj* dans la proposition subordonnée. Pour la proposition relative avec l'extraction de l'objet, il n'y a pas de dépendance *obj* dans la proposition subordonnée. Les subordonnées sont introduites par un marqueur 的 (écriture en pinyin : *de*). Comme dans UD, les mots fonctionnels dépendent du mot plein, ainsi, le marqueur 的 est lié avec le verbe de la subordonnée par relation *mark:relcl*. Ainsi, en générale, on ne rencontrera pas en même temps $O_{bas}\boldsymbol{O_{haut}}$ et $S_{bas}\boldsymbol{O_{haut}}$ dans une construction relative.

$O_{bas}O_{haut}$ est beaucoup plus fréquent par rapport à l'anglais ou au français. Un exemple de deux dépendances *obj* disjointes est présenté à la figure 94[55].

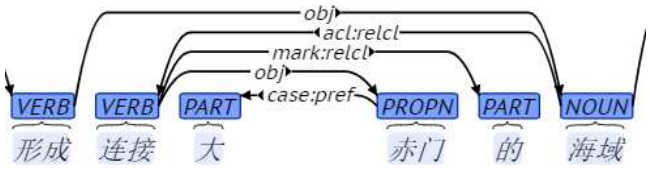

Figure 94 Exemple (UD_Chinese-GSD, train-s1645)

形成	连接	大	赤门	的	海域
Xing cheng	lian jie	da	Chimen	de	hai yu
Former	relier	grand	Chimen	PART	zone maritime

… forme une zone maritime reliée au grand Chimen …

Dans le corpus, nous avons également rencontré des cas en $\boldsymbol{O_{bas}}O_{haut}$ ou $\boldsymbol{O_{bas}}S_{haut}$, résultats des critères d'annotation (par exemple, la phrase de la figure 95). Différents treebanks ont dénommé les structures comportant les mots du type de *在 zai* (dans), ou encore *关于 guanyu* (à propos de) de manière différente. Dans le treebank GSD et le treebank PUD, ces mots sont considérés comme des verbes. Dans le treebank HK, ils sont annotés en prépositions qui se trouvent dépendantes d'une relation *case*. Pour la phrase de la figure 95, le fait de considérer ces mots comme des verbes aurait de plus comme conséquence de compter plus de cas contenant $\boldsymbol{O_{bas}}$.

[55] Il faut noter qu'en chinois, la proposition relative avec l'extraction du sujet ne contient pas de dépendance *nsubj*. Ainsi, nous ne pouvons avoir que la configuration $O_{bas}O_{haut}$ pour la construction relative de cette phrase.

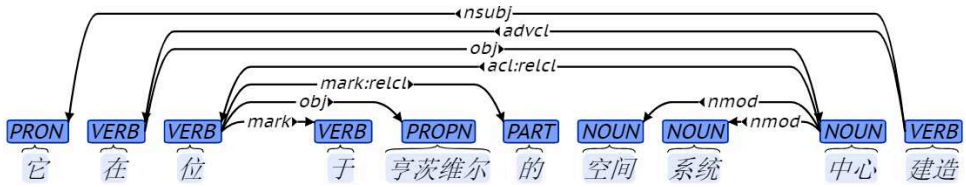

Figure 95 Exemple (UD_Chinese-GSD, test-s22)

它	在	位	于	亨茨维尔	的	空间	系统	中心	建造
Ta	zai	wei	yu	Huntsville	de	kong jian	xi tong	zhong xin	jian zao
Il	PART	se situer	à	Huntsville	PART	espace	système	centre	construire

Il a été construit au Centre des systèmes spatiaux de Huntsville

Nous avons constaté que les exemples contenant $S_{bas}S_{haut}$ sont rares. Comme dans la figure 96. Dans cet exemple, il s'agit d'une structure copule. Le nom propre *Geronimo* est le gouverneur du sujet *歌曲 gequ* (chanson).

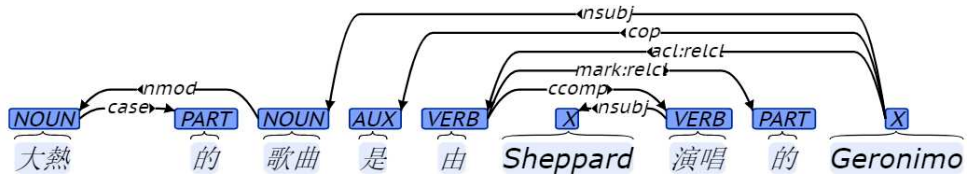

Figure 96 Exemple (UD_Chinese-PUD, n04003043)

大热	的	歌曲	是	由	Sheppard	演唱	的	Geronimo
Da re	de	ge qu	shi	you	Sheppard	yan chang	de	Geronimo
Populaire	PART	chanson	être	par	Sheppard	chanter	PART	Geronimo

La chanson à succès est Geronimo, chantée par Sheppard.

Pour conclure, d'après les résultats sur les dépendances disjointes que nous avons observés, nous n'avons finalement pas trouvé que les dépendances disjointes de même type sont beaucoup moins fréquentes que les autres. Pour les dépendances nsubj et obj, c'est la position de la proposition subordonnée de la langue qui détermine la distribution des quatre conditions que nous avons choisi d'examiner. Comme mentionné ci-dessus, l'anglais et le français ont des ordres similaires pour les mots dans la phrase, donc les distributions des configurations sont similaires : pour ces deux langues, nous avons constaté que les dépendances disjointes ayant S_{haut} sont plus fréquentes. Pour le chinois par contre, ce sont les cas en $\boldsymbol{O_{bas}}O_{haut}$ et $\boldsymbol{O_{bas}}S_{haut}$ qui sont les plus

fréquents. Nous observons là un point intéressant en ce qui concerne l'ordre des mots dans une langue, pour expliquer éventuellement les différences entre les mesures par des métriques qui prennent en compte des configurations disjointes. Dans nos futurs travaux, nous étudierons d'autres types de dépendances, ainsi que davantage de langues.

5.5.2 Poids combiné

Dans la section 4.3.3 du chapitre 4, nous avons expliqué le poids combiné qui prend en compte le nombre maximum de dépendances disjointes et les longueurs de ces dépendances dans une position de flux. Cette section se concentre sur l'étude quantitative du poids combiné dans 37 treebanks (37 langues) réalisée par Yan et Kahane (2018).

Le calcul du poids combiné se fonde sur les arbres agrégés de ces 37 treebanks. Ces arbres agrégés sont les mêmes que ceux qui ont été utilisés pour le calcul du poids (voir section 5.3.3). Le tableau 32 montre les résultats pour le poids combiné dans les arbres agrégés. Pour le poids maximum dans les 37 langues, on constate une grande diversité entre les langues : il est entre 132 (pour le galicien) et 22 (pour le vietnamien). Ainsi, il est difficile de trouver une valeur contrainte commune pour cette métrique. En ce qui concerne le poids combiné moyen dans les 37 langues, il varie entre 9,42 (pour le persan) et 2,76 (pour le slovaque).

	Poids combiné (Arbres agrégés)	
Langue	**Max**	**Moyenne**
Afrikaans	77	6,17
Arabe	56	3,67
Basque	113	7,87
Bulgare	77	4,42
Catalan	52	3,43

Chinois	54	5,21
Croate	41	7,78
Tchèque	44	4,17
Danois	62	4,55
Néerlandais	81	5,26
Anglais	53	3,96
Estonien	58	4,11
Finnois	50	4,11
Français	53	5,07
Galicien	132	4,85
Grec	39	3,83
Hébreu	39	7,16
Hindi	64	6,26
Hongrois	77	4,42
Italien	69	3,98
Letton	129	4,98
Norvégien-Bokmaal	36	3,48
Norvégien-Nynorsk	32	3,51
Persan	103	9,42
Polonais	68	2,79
Portugais	34	4,71
Roumain	38	4,42
Russe	41	3,91
Serbe	69	4,15
Slovaque	35	2,76
Slovène	44	4,03
Espagnol	45	4,22

Suédois	35	4,69
Turc	28	5,74
Ukrainien	23	4,37
Urdu	24	8,13
Vietnamien	22	3,35

Tableau 32. Poids combiné maximal et poids combiné moyen pour les arbres agrégés de 37 langues

La figure 97 montre une comparaison entre le poids combiné moyen *combined weight (avr)* et le poids moyen *weight (avr)*. On peut constater que les poids moyens des arbres agrégés sont peu dispersés en valeur et plus homogène, car ils ne prennent en compte que les niveaux d'auto-enchâssement. Le poids combiné moyen des arbres agrégés accentue les différences entre les treebanks. Comme les valeurs des longueurs de dépendances varient beaucoup selon les langues (Liu, 2008), la différence pour le poids combiné est amplifiée. D'ailleurs, la longueur de dépendance tient compte des informations linéaires, elle est corrélée avec la longueur de la phrase (Jiang & Liu, 2015). Cette dernière est sensible au genre de corpus. Comme l'information sur le genre est manquante et que l'on ne dispose que d'un treebank pour chaque langue, on ne peut pas déterminer si cette différence est due aux langues ou aux genres.

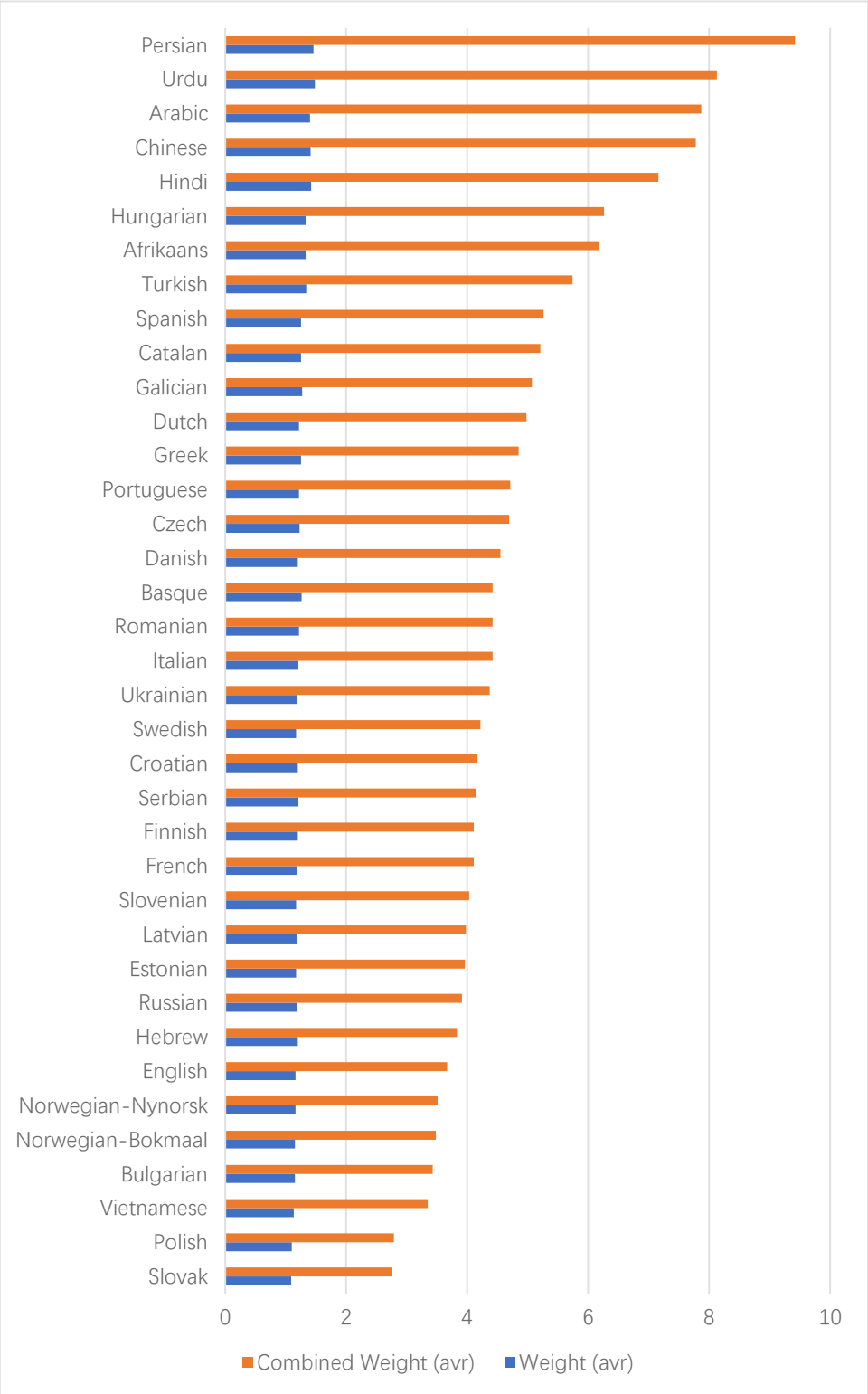

Figure 97 Poids combiné moyen (*combined weight* (*avr*)) et poids moyen (*weight* (*avr*)) dans 37 langues

5.5.3 Flux requis

Dans cette section, nous présentons nos études quantitatives du flux requis (voir 4.3.4). Nous rappelons que le flux requis d'une position représente l'ensemble des dépendances « ouvertes » à cette position lorsque la phrase est traitée de gauche à droite. Nous rappelons également que les dépendances « ouvertes » se composent de deux catégories principales : les dépendances vers la gauche pour lesquelles le dépendant déjà traité attend son gouverneur à venir ; et les dépendances vers la droite du type complément (les relations *comp* dans les treebanks SUD).

Nous rappelons aussi les différences entre le flux requis et le flux potentiel : l'élaboration du flux potentiel se fonde sur le respect de la projectivité. Alors que la composition du flux requis se fonde sur des analyses des dépendances impliquées à chaque position du flux et que les données stockées dans la mémoire de travail sont un ensemble des dépendances, le flux potentiel, lui, suppose que les données stockées dans la mémoire de travail sont un ensemble de mots.

Puisque le flux requis d'une position représente l'ensemble des dépendances indispensables à compléter, nous pensons que sa taille pourrait être moins grande que celle du flux observé. En effet, il existe des dépendances dans le flux observé qui ne font pas partie du flux requis, par exemple, celles du type modifieur vers la droite. Ainsi, en général, nous nous attendons à ce que les tailles du flux requis soient plus faibles que celles du flux observé dans les treebanks.

En outre, comme pour le cas du flux potentiel, nous pensons que la distribution des tailles du flux requis peut également refléter les différences entre les types de langues. La deuxième partie de cette section concerne les études quantitatives du flux requis par langue. Pour les langues qui sont strictement à tête finale, nous pourrons déduire que le flux requis est équivalent au flux observé : puisque le dépendant précède toujours son gouverneur, le gouverneur n'est pas encore présent lors du traitement de son dépendant. Cela crée une dépendance « ouverte » vers la gauche qui fait partie du

flux requis.

Alors que pour les langues strictement à tête initiale, puisque le gouverneur apparaît en premier, il s'agit toujours de dépendances vers la droite. Pour établir le flux requis pendant le traitement de gauche à droite, nous ne pouvons considérer que les dépendances qui lient un complément postposé. Il y a aura de plus d'autres dépendances qui ne feront pas partie du flux requis. Par conséquent, prédire le flux observé à partir du flux requis n'est pas possible.

Flux requis et flux observé dans l'ensemble des treebanks SUD

Pour tester notre hypothèse mentionnée ci-dessus. Nous avons d'abord calculé la taille du flux requis dans l'ensemble des treebanks SUD 2.7. La figure 98 et la figure 99 nous montrent la distribution de la taille du flux requis et la distribution de la taille du flux observé pour l'ensemble des treebanks SUD[56].

Nous pouvons tout d'abord constater que les positions les plus nombreuses du flux requis ont une taille égale à 1[57], soit 36,85%. Ensuite, de moins en moins de positions ont une valeur plus grande. Enfin, 99,9% des positions du flux requis ont une taille inférieure ou égale à 8 (voir 8.2.4 dans Annexes pour le tableau des résultats en détail).

Ensuite, nous pouvons remarquer que les tailles du flux observé sont généralement plus importantes que celles des flux requis dans l'ensemble de SUD. 99,9% des positions du flux observé sont de taille inférieure ou égale à 10 contre 8 pour le flux requis. Cela se

[56] Des résultats détaillés sur la distribution des tailles du flux requis et sur la distribution des tailles du flux observé sont fournies à 8.2.4 et 8.2.1.

[57] Nous pouvons constater qu'il existait de nombreux cas où le flux requis était 0 (18,5 %). Nous savons que le flux requis est construit sur un arbre partiel, et qu'il contient des dépendances qui devront être établies à chaque position inter-mot. Lorsqu'il s'agit des dépendances qui ne sont pas requises à construire, elles ne feront pas partie du flux requis. Ce qui n'est pas le cas pour le flux observé, il n'existe pas de position inter-mot du flux observé où aucune dépendance ne soit présente.

révèle conforme à nos analyses de départ (voir 8.2.1 dans Annexes pour le tableau des résultats en détail).

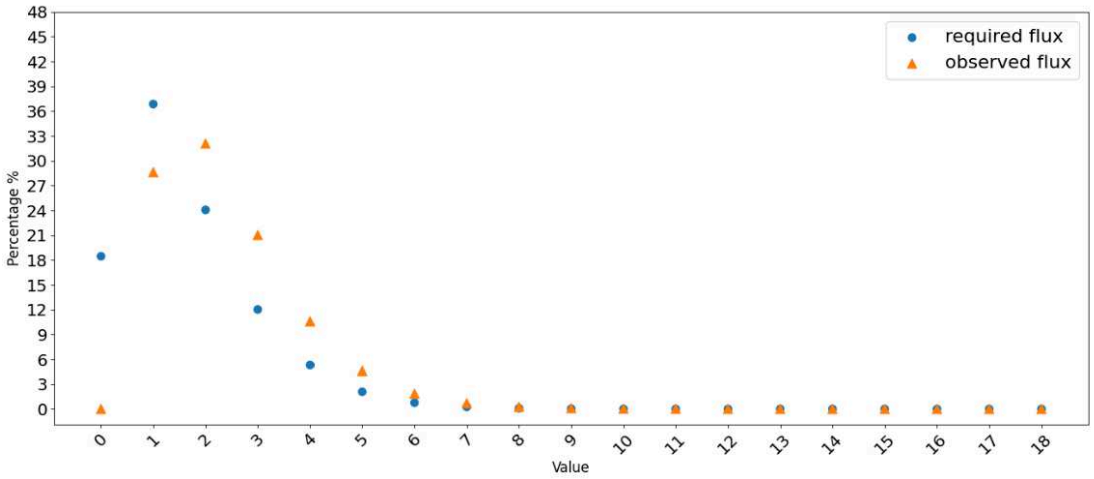

Figure 98 Distribution des tailles du flux requis et des tailles du flux observé dans SUD 2.7

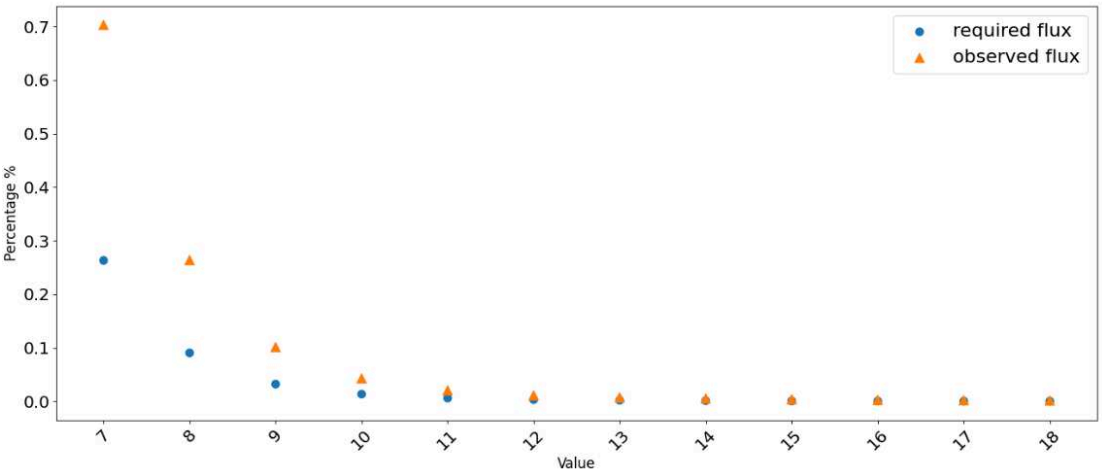

Figure 99 Distribution des valeurs entre 7 et 18 pour le flux requis et le flux observé dans SUD 2.7

Les tailles du flux requis dans 5 langues

Nous avons choisi cinq langues pour voir plus en détail la situation. Ces 5 langues sont identiques à celles des études quantitatives du flux potentiel dans la section 5.4.2. Il s'agit des trois langues à tête finale : le turc, l'hindi et le japonais, et des deux langues à tête initiale : l'arabe et l'irlandais.

La figure 100 et la figure 101 montrent la répartition des tailles du flux requis pour les cinq langues. Comme ce que nous avons analysé au début de cette section, les tailles du flux requis dans les langues à tête finale sont plus importantes que celles pour les langues à tête initiale.

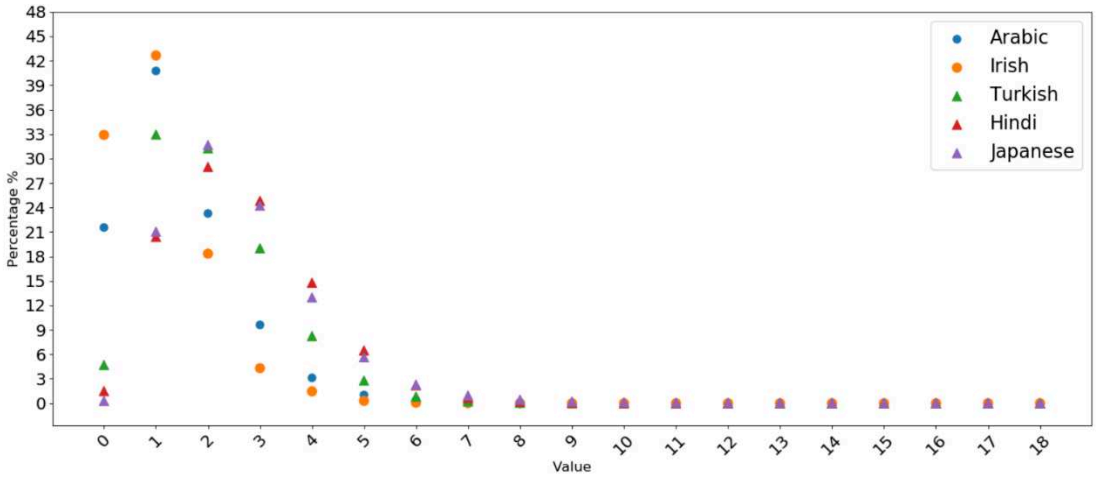

Figure 100 Flux requis pour cinq langues de SUD 2.7

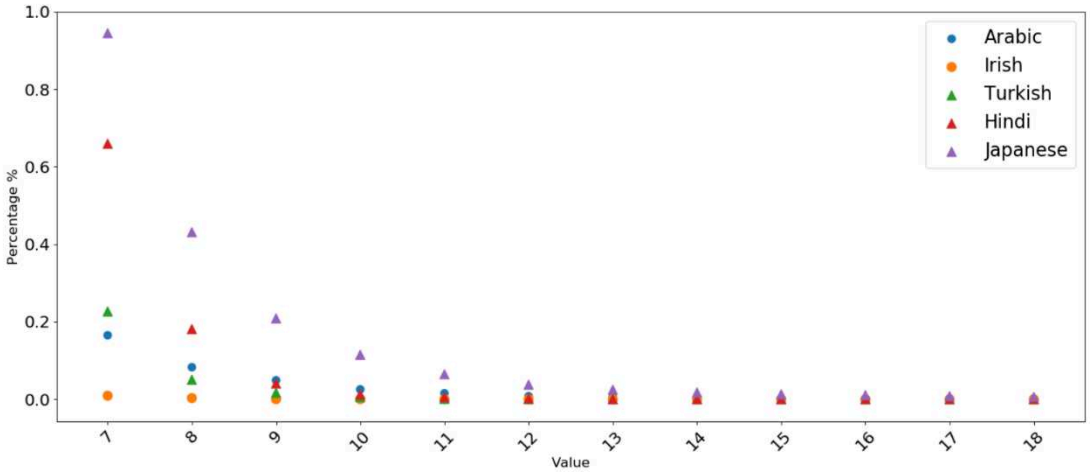

Figure 101 Flux requis pour cinq langues (valeurs 7 entre 18) de SUD 2.7

Pour conclure, en comparant la distribution des tailles du flux requis avec celle des tailles du flux observé pour ces cinq langues (voir la figure 85 et la figure 86 dans 5.4.2), nous pouvons faire les remarques suivantes :

- Pour les langues à tête finale, la répartition des tailles du flux requis est similaire à celle des tailles du flux observé, ceci est bien le cas du turc, de l'hindi et du japonais (voir aussi 8.2.5 pour plus de détails).

- En revanche, le flux observé est beaucoup plus important que le flux requis pour les langues à tête initiale.

Commentaires

Nous pouvons résumer tous nos résultats quantitatifs par les conclusions suivantes, qui, nous allons le voir, nous permettent de confirmer nos hypothèses de départ :

- La taille du flux requis est plus petite que celle du flux observé. Nous constatons que 99,9% des positions du flux requis ont une taille inférieure ou égale à 8, alors que cette valeur est de 10 pour le flux observé. Ceci est dû au fait que le flux requis n'est composé que de l'ensemble des dépendances à compléter obligatoirement pendant le traitement de gauche à droite.
- Comme le flux potentiel, les distributions des tailles du flux requis diffèrent selon les types de langue. Nous avons constaté que les langues à tête finale ont généralement des valeurs plus importantes que celles pour les langues à tête initiale. Comme le flux potentiel, les résultats du flux requis permettent également d'expliquer pourquoi la taille du flux observé dans les langues à tête initiale est plus contrainte que celle pour les langues à tête finale : le flux requis dans les langues à tête initiale fournit très peu d'informations pour établir la structure finale pendant le traitement, et donc il est difficile de prévoir le flux observé à partir du flux requis pour les langues à tête initiale. Considérant que l'incertitude entraîne une difficulté de traitement plus grande, il serait plus compliqué de rencontrer des structures complexes pour les langues à tête initiale que pour les autres.

Le flux potentiel projectif et le flux requis sont deux métriques pour la complexité du traitement syntaxique de gauche à droite, leurs résultats nous font penser que la complexité du traitement syntaxique de gauche à droite pourrait donner des contraintes sur la complexité de la structure syntaxique. D'après nos expériences, nous trouvons que les deux métriques permettent d'expliquer l'asymétrie dans les résultats du flux

observé en fonction du type de langue. Dans les travaux futurs, nous pourrons réaliser des études plus fines.

5.6 Conclusion

Dans ce chapitre, nous avons présenté nos études quantitatives pour différentes métriques du flux. Certaines valeurs des métriques sont contraintes, comme la taille du flux, le flux potentiel, ainsi que le flux requis. Pour elles, les valeurs ont tendance à être minimisées dans le processus de traitement et d'autre part on ne peut découvrir une véritable une borne supérieure expérimentalement dans les langues. Nous trouvons 99,9% de tailles du flux est inférieur à 12 dans UD et 10 dans SUD, pour le flux potentiel cette valeur est 20 dans SUD, et pour le flux requis cette valeur est de 8 dans SUD. Nous pensons que les langues naturelles ont tendance à minimiser les informations concomitantes, cette minimisation pourrait être révélée non seulement par le flux observé fondé sur l'arbre complet mais aussi le flux potentiel et le flux requis développé sur l'arbre partiel. D'ailleurs, les études dans différents types de langues pour ces métriques permettent d'expliquer que les métriques du flux potentiel et du flux requis entraînent des contraintes sur le flux observé et ceci en fonction du type de langue. Pour les langues à tête initiale, le flux potentiel et le flux requis sont tous deux difficiles à prédire : ce phénomène a pour conséquence qu'il y a plus de contraintes pour la structure d'arbre. Par conséquent, les tailles du flux observé sont plus petites dans ces langues.

En ce qui concerne le poids du flux, ses valeurs sont bornées. Nous constatons que le poids est limité à environ 5 dans les arbres originaux d'UD, et à 4 pour les arbres agrégés. Puisque le poids du flux capture le niveau d'auto-enchâssement des constructions, la limitation du poids signifie qu'il y a une limitation pour augmenter le niveau d'auto-enchâssement dans les langues naturelles. D'un point de vue cognitif, nous pensons que cela est dû à la limitation de la mémoire de travail lorsqu'elle traite

ce type de structure. Enfin, cette limitation varie en fonction de la langue écrite ou parlée, nous trouvons des valeurs plus faibles dans la langue parlée.

6 Conclusions générales et perspectives

Nos études quantitatives couvrent plus de 100 langues et certaines des expériences ont montré des résultats homogènes, ce qui nous permet d'avoir des conclusions universelles. Dans cette dernière section, nous expliquerons d'abord dans la section 6.1 comment les résultats des métriques fondées sur le flux de dépendance vérifient des hypothèses de la complexité syntaxique. Ensuite, dans la section 6.2 nous proposerons des perspectives de travaux futurs.

6.1 Métriques du flux de dépendance et complexité syntaxique

La distribution des valeurs des métriques dans les treebanks en dépendance UD/SUD

Les résultats expérimentaux des métriques basées sur le flux de dépendance montrent que leurs distributions rendent compte de la complexité syntaxique. En observant les distributions des valeurs des métriques du flux tel que le poids du flux, la taille du flux et le flux potentiel, nous avons constaté que les petites valeurs sont très fréquentes, puis plus la valeur est grande, plus elle est rare.

Plus précisément, dans les résultats pour la taille du flux et le flux potentiel, la fréquence de la distribution des valeurs diminue rapidement après avoir été supérieure à 2. De plus, dans les résultats pour le poids du flux, plus le poids est élevé, moins il est fréquent, avec une limite de poids d'environ 5 pour toutes les langues.

Si nous considérons que les phrases moins complexes syntaxiquement devraient avoir des valeurs de complexité syntaxique plus faibles, les petites valeurs de complexité syntaxique sont plus fréquentes dans les treebanks. De ce fait, nous pensons que ces

résultats de distribution sont cohérents avec notre hypothèse initiale selon laquelle les structures de phrases trop complexes devraient être rares, voire inexistantes, dans les données réelles.

Lien avec la Théorie de minimisation de longueur de dépendance

Une étude importante dans cette thèse porte sur la distribution des valeurs pour la taille du flux et pour la longueur de dépendance dans la section 5.2.

Au chapitre 4, nous avons expliqué que la somme de la longueur de dépendance et de celle de la taille du flux sont équivalentes dans la phrase. Cela nous a conduit à constater que la somme des longueurs de dépendance et la somme des tailles du flux dans plus de 100 treebanks étaient identiques et que, naturellement, leurs moyennes étaient la même. Pourtant, leurs distributions diffèrent, bien que les longueurs de dépendance et les tailles du flux tendent à être minimisées dans le corpus de 100 langues. La tendance à la minimisation de la longueur de dépendance a été observée dans de nombreuses études. Nous constatons que la distribution des tailles de flux est plus contrainte que la longueur de dépendance.

Nous rappelons ici qu'elles peuvent être interprétées selon différentes hypothèses cognitives (voir 5.2.3). L'interprétation de la minimisation de la longueur de dépendance est que nous avons tendance à réduire le temps pendant lequel les éléments sont placés dans la mémoire de travail. La distribution de la taille du flux peut donner une autre interprétation sur la théorie de la minimisation de la longueur de dépendance. C'est que l'on tend à minimiser l'information syntaxique qui est traitée simultanément.

Limites universelles de la complexité syntaxique

Pour le poids du flux, nous pensons que les résultats obtenus sont plus intéressants que la taille du flux, qui est le centre de cette thèse. Nous rappelons que le poids du flux peut représenter le niveau d'auto-enchâssement de la structure. Des limites au niveau d'auto-enchâssement ont été trouvées dans des études psycholinguistiques (Lewis,

1996). Nos observations sur la distribution des poids visent donc à trouver la valeur de poids maximale dans les treebanks UD.

Kahane, Yan et Botalla (2017) ont révélé une limitation du poids du flux dans 50 langues (70 treebanks). Cette limite est de 5. Nous avons fait un raffinement supplémentaire de leurs expériences dans la section 5.3. Parmi les treebanks en 47 langues qui ne considèrent que des dépendances entre des mots pleins, nous avons constaté que la limite du poids du flux de la plupart des langues était de 4. Cette limite universelle impliquerait une limite de la complexité syntaxique dans les langues, ce qui est probablement due à la limitation de la mémoire de travail. Cette dernière se situe également autour de 4 éléments selon les expériences psychologiques de Cowan (2001).

Pour la taille du flux ainsi que pour d'autres métriques, nous n'avons pas trouvé de valeur qui pourrait être une limite universelle. Par exemple, les calculs effectués pour les 49 treebanks UD au chapitre 5 ont montré que les valeurs contraintes de la taille du flux que nous avons observées allaient de 7 à 32, avec une valeur moyenne en 11,30. La taille du flux n'est pas aussi contrainte que le poids du flux, mais plus contrainte que la longueur de dépendance, qui donne des valeurs contraintes entre 11 et 75 pour les mêmes treebanks, avec une moyenne de 32,64. Nous signalons que l'effet du schéma d'annotation sur nos résultats ne peut pas non plus être ignoré. En UD, les éléments en conjonction de coordination sont annotés en bouquet, par conséquent, un grand nombre de conjoints se traduira par une taille de flux très élevée.

Complexité syntaxique dans différents types de treebanks

Nous avons étudié dans la section 5.3.2 les valeurs du poids de flux dans les treebanks du français parlé et écrit. Nous avons constaté que les poids du flux dans le treebank du français parlé sont généralement plus faibles par rapport aux treebanks du français écrit. Nous rappelons que pour le treebank en français parlé, le poids maximal est de 4, alors que dans les trois autres treebanks en français écrit, le poids maximal est de 5. De plus, le treebank en français parlé contient des poids de flux à 1 beaucoup plus nombreux

que dans le français écrit, soit 70,54%, contre 64,31%, 60,31% et 61,25% dans les trois autres en français écrit.

Dans le contexte du traitement des phrases, des phrases produites spontanément à l'oral conduisent au fait qu'il est difficile de retenir toutes les informations nécessaires au cours de l'audition. Par contre, pour une lecture, il est possible de faire la régression lorsque le lecteur a rencontré une difficulté de compréhension. Dans le contexte de la production de phrases, lors d'un discours oral, il faut une plus grande charge sur la mémoire de travail pour générer rapidement des phrases sans aucune préparation. Nous soutenons qu'avec la préparation, les locuteurs ont tendance à se rappeler plus facilement les éléments qui doivent être utilisés, et que l'activation de ces éléments était déjà renforcée pendant la préparation. En revanche, sans préparation, il devient plus difficile de faire appel temporairement à des éléments moins activés.

La complexité syntaxique est également influencée par le type des langues étudiées. Dans la section 5.2.2, nous avons constaté que la taille du flux des langues à tête initiale est plus petite en moyenne que celle des langues à tête finale. En revanche, dans la section 5.4.2, nous avons constaté que la taille du flux potentiel est plus importante dans les langues à tête initiale par rapport aux langues à tête finale. Nous rappelons ici la différence entre le flux observé et le flux potentiel (projectif) : le flux observé pour une position se base sur l'arbre complet et capture toutes les dépendances qui passe par cette position ; le flux potentiel pour une position se base sur l'arbre partiel. Il contient des mots à gauche de cette position, qui peuvent avoir une dépendance avec des mots à droite de cette position, tout en respectant le principe de la projectivité. Dans une langue à tête initiale typique, le gouverneur d'une dépendance précède toujours son dépendant. De ce fait, chaque mot qui apparait après un gouverneur peut être son dépendant sans que cela n'entraine la non-projectivité. C'est la raison pour laquelle nous observons de plus grandes valeurs du flux potentiel dans les treebanks des langues à tête initiale. De plus, puisque le flux potentiel contient des informations nécessaires pour prédire les structures (projectives) possibles, il devient plus difficile de prédire la structure en cas

des tailles du flux potentiel importantes. Cela explique la taille relativement plus petite du flux observé dans les langues à tête initiale, parce que les structures dans ces langues sont plus contraintes en termes de complexité syntaxique.

6.2 Perspectives

Nos expériences sur des métriques basées sur le flux dans le chapitre 5 ont obtenu des résultats satisfaisants, notamment en ce qui concerne la distribution des poids du flux. Nous avons été conscients également de difficultés dans l'étude de la complexité syntaxique et des limites de notre étude. Cela constituera l'objet de recherche pour de futurs travaux.

Futurs travaux sur les métriques du flux

Tout d'abord, nous soutenons que les études quantitatives de la complexité syntaxique ne peuvent se limiter à celles de la longueur de dépendance et au cadre de la théorie de la minimisation de la longueur de dépendance. En d'autres termes, en plus de considérer la taille du flux ou la longueur de dépendance, il faut également prendre en compte les propriétés structurelles : par exemple, la prise en compte de la projectivité pour avoir le flux potentiel, ou encore la prise en compte des dépendances disjointes qui impliquent les structures auto-enchâssées afin d'avoir le poids du flux.

D'ailleurs, nous avons proposé dans la section 4.3.2.2, une formule de complexité syntaxique C (voir formule F6) qui est une combinaison linéaire du poids du flux et de la taille du flux, ce qui peut être considéré comme l'un des sujets de recherche dans le futur. Comme la formulation contient un paramètre a, à l'aide des données psycholinguistiques sur la complexité syntaxique (voir aussi plus loin : **Evaluer nos métriques à l'aide d'expériences psycholinguistiques**), nous pouvons calculer C dans des treebanks et ensuite faire des analyses.

Futures études sur les dépendances disjointes

Pour l'étude des dépendances disjointes dans 5.5.1, seule la distribution des relations *nsubj* et *obj* est prise en compte et seules trois langues (anglais, français et chinois) sont considérées. Ainsi, nous devons encore faire beaucoup de travail. Par exemple, les futurs travaux peuvent observer plus de langues et plus de configurations de dépendances.

Evaluer nos métriques à l'aide d'expériences psycholinguistiques

En analysant quantitativement la distribution des valeurs des métriques du flux dans les treebanks UD / SUD, nous soutenons qu'elles mesurent toutes des aspects de la complexité syntaxique. Pourtant, en raison du manque de comparaison des données issues d'expériences psycholinguistiques, nous ne sommes pas sûrs de savoir lesquelles sont les plus proches des données réelles de complexité syntaxique. Les prochaines pistes de travaux consistent à effectuer des études psycholinguistiques afin d'avoir des données réelles sur la complexité syntaxique. Dans ce processus, une étude qualitative est tout d'abord nécessaire. Nous devons trouver des phrases dans le corpus qui peuvent être utilisées dans les expériences basées sur différentes hypothèses. Ensuite, nous devrons choisir une bonne méthode pour évaluer la complexité syntaxique de ces phrases. Dans la section 4.4 du chapitre 4, nous avons déjà abordé les avantages et les inconvénients des méthodes possibles. Nous pensons enfin qu'il est plus approprié d'appliquer les expériences en utilisant la méthode d'oculométrie.

7 Bibliographie

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a treebank for French. In *Treebanks* (pp. 165-187). Springer, Dordrecht.

Abney, S. P. (1989). A computational model of human parsing. *Journal of psycholinguistic Research*, *18*(1), 129-144.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341-380.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Baddeley, A., Logie, R., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluent reading. *Journal of memory and language*, *24*(1), 119-131.

Beauzée, N. (1765). Régime, in Denis Diderot & Jean Le Rond D'Alembert J. (eds.), *Encyclopédie*, vol.

14, 5-11.

Benzitoun, C., Dister, A., Gerdes, K., Kahane, S., Pietrandrea, P., Sabio, F., & Debaisieux, J. M. (2010). Tu veux couper là faut dire pourquoi. Propositions pour une segmentation syntaxique du français parlé. *2ème Congrès Mondial de Linguistique Française*, 139.

Bever, T. G. (1970). *The influence of speech performance on linguistic structures*. Na.

Blache, P. (2011). A computational model for linguistic complexity. In *Biology, Computation and Linguistics* (pp. 155-167). IOS Press.

Blanche-Benveniste, C., Bilger, M., Rouget, C., Van Den Eynde, K., Mertens, P., & Willems, D. (1990). Le français parlé(études grammaticales). *Sciences du langage*.

Blanche-Benveniste, C. (2010). *Approches de la langue parlée*. Éditions Ophrys.

Bloomfield, L. (1933). Language (New York: Henry Holt, 1933). *Linguistic Aspects of Science*, 20-45.

Bodirsky, M., Kuhlmann, M., & Möhl, M. (2009). Well-nested drawings as models of syntactic structure. In *Tenth Conference on Formal Grammar and Ninth Meeting on Mathematics of Language* (pp. 195-203).

Böhmová, A., Hajič, J., Hajičová, E., & Hladká, B. (2003). The Prague dependency treebank. In *Treebanks* (pp. 103-127). Springer, Dordrecht.

Botalla, M. A. (2014). *Analyse du flux de dépendance dans un corpus de français oral annoté en microsyntaxe* (Doctoral dissertation, Master thesis. Université Sorbonne Nouvelle).

Booth, T. L. (1969). Probabilistic representation of formal languages. In *10th annual symposium on switching and Automata Theory (swat 1969)* (pp. 74-81). IEEE.

Bradley D. C., Garrett M. E., Zurif E. B. (1980). Syntactic deficits in Broca's aphasia. In Biological Studies of Mental Processes, Caplan D., editor. ed. (Cambridge, MA, MIT Press;).

Brants, T., Hendriks, R., Kramp, S., Krenn, B., Preis, C., Skut, W., & Uszkoreit, H. (1999). NEGRA Annotierschema. *unpublished, Arbeitsmaterial*, *20*.

Buffier, C. (1709). *Grammaire française sur un plan nouveau* [French grammar on a new plan], Paris: Le Clerc- Brunet-Leconte & Montalant.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and brain Sciences*, *22*(1), 77-94.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. *Spoken and written language: Exploring orality and literacy*, 35-54.

Chomsky, N. (1957). *Syntactic structure*. Mouton.

Chomsky, N. (1965). Aspects of the theory of syntax (Vol. 11). *MIT Press. doi*, *10*, 90008-5.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59*, 126-161.

Clifton Jr, C., & Staub, A. (2011). Syntactic influences on eye movements during reading. *eye*, *3*(2).

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational linguistics*, *29*(4), 589-637.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, *56*(13), 1448-1462.

Courtin, M., & Yan, C. (2019). What can we learn from natural and artificial dependency trees.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why. *Current directions in psychological science*, *19*(1), 51-57.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. Educational research bulletin, 37-54.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain research*, *1084*(1), 89-103.

Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 561.

Deulofeu, J. (2003). L'approche macrosyntaxique en syntaxe: un nouveau modèle de rasoir d'Occam contre les notions inutiles. *Scolia*, *16*, 77-95.

De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014, May). Universal Stanford dependencies: A cross-linguistic typology. In *LREC* (Vol. 14, pp. 4585-4592).

Duval, C., Piolino, P., Bejanin, A., Laisney, M., Eustache, F., & Desgranges, B. (2011). La théorie de l'esprit: aspects conceptuels, évaluation et effets de l'âge. *Revue de neuropsychologie*, *3*(1), 41-51.

Embick, D., Marantz, A., Miyashita, Y., O'Neil, W., & Sakai, K. L. (2000). A syntactic specialization for Broca's area. *Proceedings of the National Academy of Sciences*, *97*(11), 6150-6154.

Ferre-i-Cancho, R. (2006). Why do syntactic links not cross?. *EPL (Europhysics Letters)*, *76*(6), 1228.

Flesch, R. (1948). A new readability yardstick. Journal of applied psychology, 32(3), 221.

Frazier, L. (1979). *On comprehending sentences: syntactic parsing strategies. 1979* (Doctoral dissertation, Doctoral dissertation-University of Connecticut, Distributed by Indiana Linguistics Club).

Fodor, J. A., Garrett, M., & Bever, T. G. (1968). Some syntactic determinants of sentential complexity, II: Verb structure. *Perception & Psychophysics*, *3*(6), 453-461.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & psychophysics*, *8*(4), 215-221.

Friederici, A. D. (1995). The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and language*, *50*(3), 259-281.

Friederici, A. D., Hahne, A., & Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. *Journal of psycholinguistic research*, *31*(1), 45-63.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, *6*(2), 78-84.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, *91*(4), 1357-1392.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336-10341.

Gerdes, K., & Kahane, S. (2017). Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral: le cas de la macrosyntaxe.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018, November). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD.

Gerdes, K., Kahane, S., & Chen, X. (2021). Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*, *6*(1).

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1-76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, *2000*, 95-126.

Gibson, E., & Warren, T. (2004). Reading-Time Evidence for Intermediate Linguistic Structure in Long-Distance Dependencies. *Syntax*, *7*(1), 55-78.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length?. *Cognitive Science*, *34*(2), 286-310.

Gómez-Rodríguez, C., & Ferrer-i-Cancho, R. (2017). Scarcity of crossing dependencies: A direct outcome of a specific constraint?. *Physical Review E*, *96*(6), 062304.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, 25(2), 149188.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *In J. Greenberg, ed., Universals of Language. 73-113. Cambridge, MA.*

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, *29*(2), 261-290.

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of cognitive neuroscience*, *15*(6), 883-899.

Hajic, J., Vidová-Hladká, B., & Pajas, P. (2001, December). The prague dependency treebank: Annotation structure and support. In *Proceedings of the IRCS Workshop on Linguistic Databases* (pp. 105-114).

Hale, J. (2001, June). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics.

Hale, J. (2003). The information conveyed by words in sentences. *J. Psychol. Res*. 32, 101–123.

Havelka, J. (2007). Mathematical Properties of Dependency Trees and their Application to Natural Language Syntax.

Hawkins, J. A. (1994). *A performance theory of order and constituency* (Vol. 73). Cambridge University Press.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press on Demand.

Holmes, V. M., & Forster, K. I. (1972). Perceptual complexity and underlying sentence structure. *Journal of Verbal Learning and Verbal Behavior*, *11*(2), 148-156.

Holmes, V.M. (1973). Order of Main and Subordinate Clauses in Sentence Perception. *Journal of Verbal Learning and Verbal Behavior* 12(3): 285-93.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior*, *20*(4), 417-430.

Hudson, R. (1995). Measuring syntactic difficulty. *Manuscript, University College, London*.

Ihm, P., & Lecerf, Y. (1963). Éléments pour une grammaire générale des langues projectives.

Jardonnet, U. (2009). Analyse du flux de dépendance. Mémoire de master, Université Paris Ouest Nanterre La Défense

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

Jing, Y., & Liu, H. (2015, August). Mean Hierarchical Distance Augmenting Mean Dependency Distance. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)* (pp. 161-170).

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics*, *53*, 61-79.

Kaan, Edith, et al. "The P600 as an index of syntactic integration difficulty." *Language and cognitive processes* 15.2 (2000): 159-201.

Kahane, S. (2001). Grammaires de dépendance formelles et théorie Sens-Texte. *TALN 2001*.

Kahane, S., Yan, C., & Botalla, M. A. (2017, September). What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks.

Kahane, S., & Yan, C. (2019, August). Advantages of the flux-based interpretation of dependency length minimization.

Kahane, S., & Gerdes, K. (2021). *Syntaxe théorique et formelle, Volume 1 : Modélisation, unités, structures*. Language Science Press, 494 p.

Kemper, S., Kynette, D., Rash, S., O'Brien, K., & Sprott, R. (1989). Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics*, *10*(1), 49-66.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, *30*(5), 580-602.

Koch, P., & Oesterreicher, W. (2001). Language parlé et langage écrit, Lexikon der romanistischen Linguistik 1, 584–627. *Tübingen: Max Niemeyer Verlag*.

Köhler, R., & Altmann, G. (2000). Probability Distributions of Syntactic Units and Properties∗. *Journal of Quantitative Linguistics*, *7*(3), 189-200.

Kromann, M. T., Mikkelsen, L., & Lynge, S. K. (2003). Danish dependency treebank. *Proceedings of TLT2003. Växjö University Press, Sweden*.

Kuhlmann, M., & Nivre, J. (2006, July). Mildly non-projective dependency structures. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 507-514). Association for Computational Linguistics.

Kutas, M., & Hillyard, S. A. (1984). Event-Related Brain Potentials (ERPs) Elicited by Novel Stimuli during Sentence Processing a. *Annals of the New York Academy of Sciences*, *425*(1), 236-241.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J. P., ... & Tchobanov, A. (2014, May). Rhapsodie: a prosodic-syntactic treebank for spoken french.

Lakoff, R. (1979). Expository writing and the oral dyad as points on a communicative continuum: Writing anxiety as the result of mistranslation. *Unpublished manuscript*.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In *Sentence processing* (pp. 90-126). Psychology Press.

Lewis, R. (1996). A theory of grammatical but unacceptable embeddings. *Journal of Psycholinguistic Research*, *25*(93), 116.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, *29*(3), 375-419.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447-454.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159-191.

Lieven, E. V. (1978). *Conversations between mothers and young children: individual differences and their possible implication for the study of language learning* (pp. 173-187).

Makuuchi, M., Bahlmann, J., Anwander, A., & Friederici, A. D. (2009). Segregating the core computational faculty of human language from working memory. *Proceedings of the National Academy of Sciences*, *106*(20), 8362-8367.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Marcus, S. (1965). Sur la notion de projectivité. *Mathematical Logic Quarterly*, *11*(2), 181-192.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank.

Mel'čuk, I. A. (1988). *Dependency syntax: theory and practice*. SUNY press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users.

Murata, M., Uchimoto, K., Ma, Q., & Isahara, H. (2001, February). Magical number seven plus or minus two: Syntactic structure recognition in Japanese and English sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 43-52). Springer, Berlin, Heidelberg.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. In *Proceedings of the Eighth International Conference on Parsing Technologies* (pp. 149-160).

Nivre, J. (2003). Theory-supporting treebanks. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories* (pp. 117-128).

Nivre, J. (2008). 13. In: Lüdeling, A., & Kytö, M. (Eds.), *Corpus Linguistics. Volume 1*. Walter de Gruyter.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Tsarfaty, R. (2016, May). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659-1666).

Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., ... & Badmaeva, E. (2017). Universal Dependencies 2.1.

Nivre, J., Abrams, M., Agić, Z., Ahrenberg, L., Aleksandravičiūtė, G., & Antonsen, L. Universal Dependencies 2.4; 2019. *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University*.

Nivre, J. (s. d.). *TRANSITION-BASED PARSING*. 12.

Ochs, E. (1979). Planned and unplanned discourse. In *Discourse and syntax* (pp. 51-80). Brill.

O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech*, *49*(1/2), 102-110.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785-806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of experimental psychology: Learning, memory, and cognition*, *20*(4), 786.

Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint*

arXiv:1104.2086.

Richards, Brian J. & David Malvern. 1997. Quantifying lexical diversity in the study of language development. Reading: Faculty of Education and Community Studies.

Rogalsky, C., Matchin, W., & Hickok, G. (2008). Broca's area, sentence comprehension, and working memory: an fMRI study. *Frontiers in human neuroscience*, *2*, 14.

Rosa, R., Masek, J., Marecek, D., Popel, M., Zeman, D., & Zabokrtský, Z. (2014, May). HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *LREC* (pp. 2334-2341).

Sasaki, F., Witt, A., & Metzing, D. (2003). Declarations of relations, differences and transformations between theory-specific treebanks: A new methodology.

Sampson, G. (2002). English for the computer: The SUSANNE corpus and analytic scheme.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997). An annotation scheme for free word order languages. *arXiv preprint cmp-lg/9702004*.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition, 116,* 71-86.

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. In *Treebanks* (pp. 5-22). Springer, Dordrecht.

Temperley, D. (2008). Dependency-length minimization in natural and artificial languages∗. *Journal of Quantitative Linguistics*, *15*(3), 256-282.

Tesnière, L. (1965). Éléments de syntaxe structurale (1959). *Paris: Klincksieck*.

Tesnière, L., & Kahane, S. (2015). Elements of structural syntax. *New York: John Benjamins Publishing Company*.

Ure, J. (1971). Lexical density and register differentiation. *Applications of linguistics*, *443452*.

Wells, R. S. (1947). Immediate constituents. *Language*, *23*(2), 81-117.

Wehrli, E. (1989). Deux problèmes d'analyse syntaxique automatique. *Cahiers de LinguistiqueFrançaise*, (10),27-41

Wundt, W. M. (1900). *Völkerpsychologie: Bd., 1.-2. T. Die Sprache. 1900* (Vol. 1). W. Englemann.

Yamashita, H., & Chang, F. (2001). *Long before short preference in the production of a head-final language*. 11.

Yan, C., & Kahane, S. (2018). Syntactic complexity combining dependency length and dependency flux weight. *Proceedings of the First Shared Task on Measuring Language Complexity*, 38-43.

Yan, C. (2017). *Étude du flux de dépendance dans 70 corpus (50 langues) de UD.* (Master thesis. Université Sorbonne Nouvelle).

Yang, J. (2019). Syntactic Hierarchy Depth: Distribution, Interrelation and Cross-Linguistic Properties. *Journal of Quantitative Linguistics*, *26*(2), 129-145.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, *104*(5), 444-466.

Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., ... & Hajič, J. (2014). HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, *48*(4), 601-637.

8 Annexes

8.1 Algorithmes

Deux algorithmes de base sont nécessaires pour les expériences dans cette thèse, l'un pour calculer le poids du flux et l'autre pour calculer le flux potentiel. Nous n'entrerons pas dans les détails ici pour d'autres métriques, parce que certaines sont moins complexes à calculer, comme la taille du flux, et que certaines découlent de l'un de ces deux algorithmes, comme le poids combiné.

8.1.1 Poids du flux

On peut déterminer visuellement le poids du flux dans l'arbre de dépendance. Pour calculer les poids du flux dans un treebank, il nous faut un algorithme qui le calcule dans la matrice du flux. Kahane, Yan et Botalla (2017) ont proposé un algorithme simple en temps linéaire. Dans les travaux de Yan (2017), nous décrivons l'idée de cet algorithme de la façon suivante :

> *« Pour le calculer, nous pouvons commencer par une dépendance D dans le flux avec au moins un sommet qui n'est pas partagé avec d'autres dépendances dans le flux (une telle dépendance existe car la structure est acyclique). Ensuite, nous supprimons toutes les dépendances qui partagent un sommet avec D et ne peuvent donc pas être disjointes de D. Si le flux qui reste n'est pas vide, nous commençons par le même processus exactement : choisir une dépendance avec au moins un sommet qui n'est pas partagé avec d'autres dépendances dans le flux restant et en supprimant toutes les dépendances partageant un sommet avec lui. À la fin, nous obtenons l'un des plus grands ensembles de dépendances disjointes dans le*

flux. »

Les tableau 33 jusqu'à tableau 36 montrent le processus de calcul du poids du flux. Le tableau 33 montre l'étape 0, qui est la matrice originale du flux. Dans l'étape 1 (Tableau 34), une relation est choisie, elle est seule dans sa ligne ou seule dans sa colonne. Ici, la relation choisie est *Dep A* car, étant seule dans sa ligne, la valeur du poids est augmentée de 1. Dans l'étape 2 (Tableau 35), toutes les dépendances partageant la même ligne ou colonne de *Dep A* sont supprimées dans la matrice. Ainsi, on obtient une nouvelle matrice ayant seulement deux dépendances. On continue avec la même procédure des étapes 1 et 2 pour la nouvelle matrice, jusqu'à ce que la matrice soit vide (Tableau 36, étape finale).

	1	2	3
1	*Dep A*		
2	*Dep B*		
3		*Dep C*	*Dep D*

Tableau 33 Matrice (étape 0). Poids = 0

	1	2	3
1	*Dep A*		
2	*Dep B*		
3		*Dep C*	*Dep D*

Tableau 34 Matrice (étape 1). Poids + 1

	1	2
1	*Dep C*	*Dep D*

Tableau 35 Matrice (étape 2). Poids + 1

	1	2
1		

Tableau 36 Matrice (étape finale). Poids = 2

Cet algorithme permet de récupérer le poids du flux, mais aussi le poids combiné. Il faut juste ajouter l'information de la longueur pour chaque dépendance dans la matrice. Afin de trouver les relations disjointes ayant la longueur maximale, il faut dans l'étape 2 ajouter un processus pour comparer les relations à supprimer et pour garder la valeur la plus grande de longueur de dépendance de ces relations.

8.1.2 Flux potentiel

Avec la présentation du flux potentiel dans la section 4.3.2.1, il y a trois remarques pour nous aider à mettre en œuvre l'algorithme pour calculer les flux potentiels.

Premièrement, lorsqu'il s'agit d'une dépendance vers la gauche (voir la figure 102), seul le gouverneur fait partie du flux potentiel.

Figure 102

Deuxièmement, pour une dépendance vers la droite (voir la figure 103), le gouverneur et son dépendant font partie du flux potentiel.

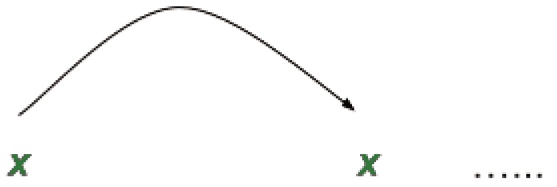

Figure 103

Enfin, les nœuds entre un gouverneur et ses dépendants qui sont déjà traités ne font pas partie du flux potentiel (voir la figure 104).

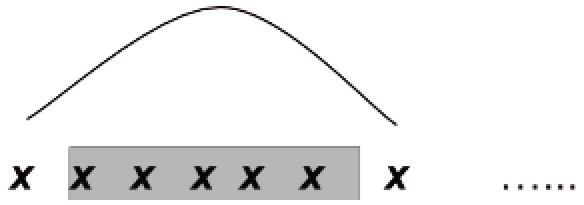

Figure 104

L'algorithme du flux potentiel est proposé ci-dessous :

Initialiser une liste vide pour le flux potentiel (*pf_list*).

Dans chaque arbre, examiner nœud par nœud de gauche à droite (nœud *X*) :

– si *X* n'a pas de gouverneur, c'est un nœud incomplet, l'ajouter dans *pf_list* ;

– si X est le gouverneur pour les éléments déjà traités, nous avons la ou les dépendances vers la gauche :

1. Ajouter *X* dans *pf_list* (si X n'est pas dans *pf_list*)

2. Supprimer les éléments dans *pf_list* qui sont entre *X* et sa dépendance la plus à gauche ;

– si *X* a déjà un gouverneur, nous avons une dépendance vers la droite :

1. Ajouter *X* et son gouverneur dans *pf_list* (s'ils ne sont pas encore dans *pf_list*)

2. Supprimer les éléments de *pf_list* qui se trouvent entre *X* et son gouverneur, à l'exception des éléments qui n'ont pas encore le gouverneur.

Le nombre d'éléments dans *pf_list* est la valeur du flux potentiel pour la position juste après *X*.

Attention : vérifiez à chaque fois que vous ajoutez un nœud dans *pf_list* : si ce nœud est supprimé auparavant dans l'historique, ne l'ajoutez pas à nouveau (pour la raison du

problème de non-projectivité).

8.2 Résultats

8.2.1 Flux observé dans SUD

n	% flux observé = n	% flux observé ≤n
0	0	0
1	28,6339	28,6339
2	32,0867	60,7206
3	21,0313	81,7519
4	10,5949	92,3469
5	4,6280	96,9748
6	1,8491	98,8240
7	0,7032	99,5272
8	0,2640	99,7912
9	0,1015	99,8927
10	0,0429	99,9356
11	0,0206	99,9562
12	0,0112	99,9674
13	0,0073	99,9747
14	0,0055	99,9802
15	0,0038	99,9840
16	0,0027	99,9867
17	0,0021	99,9888
18	0,0017	99,9905
19	0,0012	99,9917

20	0,0011	99,9928

8.2.2 Distribution des tailles du flux potentiel dans cinq langues dans SUD

n	**Arabe**		**Irlandais**	
	% flux potentiel projectif= n	% flux potentiel projectif ≤n	% flux potentiel projectif= n	% flux potentiel projectif ≤n
1	4.9716	4.9716	6.6769	6.6769
2	8.2257	13.1973	17.0274	23.7043
3	10.6579	23.8552	19.9587	43.6631
4	11.4359	35.2912	17.0113	60.6744
5	11.2374	46.5286	12.8852	73.5596
6	10.4778	57.0063	9.1516	82.7112
7	9.2316	66.2380	6.2163	88.9275
8	7.7723	74.0102	4.0701	92.9976
9	6.2854	80.2956	2.6259	95.6235
10	4.9437	85.2393	1.6434	97.2669
11	3.7891	89.0284	1.0085	98.2754
12	2.8768	91.9051	0.6039	98.8793
13	2.1115	94.0166	0.3716	99.2509
14	1.5750	95.5916	0.2524	99.5033
15	1.1717	96.7633	0.1572	99.6605
16	0.8657	97.6290	0.0991	99.7596
17	0.6284	98.2575	0.0571	99.8167
18	0.4611	98.7186	0.0351	99.8518
19	0.3430	99.0616	0.0290	99.8808
20	0.2487	99.3102	0.0270	99.9079

21	0.1779	99.4882	0.0230	99.9309
22	0.1290	99.6172	0.0170	99.9479
23	0.0992	99.7164	0.0110	99.9589
24	0.0738	99.7902	0.0090	99.9680
25	0.0541	99.8443	0.0070	99.9750
26	0.0377	99.8820	0.0050	99.9800
27	0.0283	99.9102	0.0040	99.9840
28	0.0220	99.9322	0.0020	99.9860
29	0.0145	99.9467	0.0020	99.9880
30	0.0127	99.9594	0.0020	99.9900
31	0.0091	99.9685	0.0020	99.9920
32	0.0074	99.9759	0.0030	99.9950
33	0.0055	99.9814	0.0020	99.9970
34	0.0036	99.9850	0.0030	100.0000
35	0.0025	99.9875		
	0.0026	99.9901		
	0.0027	99.9929		
	0.0022	99.9951		
	0.0010	99.9961		
	0.0009	99.9969		
	0.0005	99.9975		
	0.0005	99.9980		
	0.0008	99.9988		
	0.0003	99.9991		
	0.0005	99.9997		
	0.0003	100.0000		

n	Turc		Hindi		Japonais	
	% flux potentiel projectif= n	% flux potentiel projectif ≤n	% flux potentiel projectif= n	% flux potentiel projectif ≤n	% flux potentiel projectif= n	% flux potentiel projectif ≤n
1	27.4775	27.4775	20.3141	20.3141	25.1637	25.1637
2	27.7098	55.1873	21.6867	42.0008	30.2223	55.3860
3	20.5601	75.7474	22.0080	64.0088	22.2417	77.6277
4	12.5635	88.3108	16.5060	80.5148	12.2427	89.8704
5	6.3950	94.7058	10.3377	90.8525	5.7271	95.5975
6	2.9850	97.6908	5.2286	96.0811	2.5334	98.1308
7	1.3076	98.9984	2.3253	98.4064	1.0649	99.1958
8	0.5511	99.5495	0.9072	99.3135	0.4477	99.6435
9	0.2138	99.7633	0.3646	99.6781	0.1891	99.8326
10	0.1088	99.8721	0.1376	99.8157	0.0878	99.9205
11	0.0547	99.9268	0.0795	99.8952	0.0366	99.9571
12	0.0286	99.9555	0.0437	99.9389	0.0176	99.9747
13	0.0102	99.9656	0.0214	99.9603	0.0089	99.9835
14	0.0083	99.9739	0.0126	99.9729	0.0052	99.9887
15	0.0064	99.9803	0.0096	99.9825	0.0031	99.9918
16	0.0051	99.9854	0.0072	99.9898	0.0019	99.9938
17	0.0057	99.9911	0.0072	99.9970	0.0013	99.9950
18	0.0045	99.9955	0.0021	99.9991	0.0012	99.9962
19	0.0045	100.0000	0.0006	99.9997	0.0008	99.9970
20			0.0003	100.0000	0.0008	99.9978
21					0.0006	99.9985
22					0.0005	99.9990

23					0.0004	99.9994
24					0.0002	99.9996
25					0.0002	99.9998
26					0.0002	99.9999
27					0.0001	100.0000

8.2.3 Flux observé dans les cinq langues de SUD

n	Arabe		Irlandais	
	% flux observé = n	% flux observé ≤ n	% flux observé = n	% flux observé ≤ n
0	0	0	0	0
1	26,4162	26,4162	24,0508	24,0508
2	32,3137	58,7299	35,3687	59,4195
3	21,8067	80,5367	23,4269	82,8464
4	10,7852	91,3218	10,8552	93,7016
5	4,6125	95,9343	3,9759	97,6775
6	1,9405	97,8749	1,3200	98,9975
7	0,8632	98,7381	0,4497	99,4472
8	0,4265	99,1646	0,1863	99,6335
9	0,2418	99,4064	0,1132	99,7466
10	0,1507	99,5571	0,0801	99,8267
11	0,1024	99,6594	0,0511	99,8778
12	0,0730	99,7325	0,0391	99,9169
13	0,0572	99,7896	0,0290	99,9459
14	0,0474	99,8371	0,0160	99,9619
15	0,0323	99,8694	0,0120	99,9740
16	0,0209	99,8903	0,0160	99,9900

17	0,0147	99,9050	0,0090	99,9990
18	0,0129	99,9179	0,0010	100
19	0,0083	99,9262		
20	0,0069	99,9331		
21	0,0066	99,9397		
22	0,0045	99,9442		
23	0,0027	99,9469		
24	0,0034	99,9503		
25	0,0042	99,9544		
26	0,0031	99,9575		
27	0,0035	99,9610		
28	0,0042	99,9652		
29	0,0046	99,9698		
30	0,0030	99,9727		
31	0,0019	99,9746		
32	0,0012	99,9758		
33	0,0015	99,9773		
34	0,0011	99,9784		
35	0,0010	99,9794		
36	0,0008	99,9802		
37	0,0009	99,9811		
38	0,0005	99,9816		
39	0,0012	99,9828		
40	0,0008	99,9836		
41	0,0003	99,9839		
42	0,0005	99,9845		
43	0,0004	99,9849		

44	0,0003	99,9852		
45	0,0003	99,9855		
46	0,0003	99,9859		
47	0,0004	99,9863		
48	0,0003	99,9866		
49	0,0004	99,9871		
50	0,0004	99,9875		
51	0,0005	99,9881		
52	0,0008	99,9888		
53	0,0003	99,9892		
54	0,0005	99,9897		
55	0,0003	99,9900		
56	0,0003	99,9904		
57	0,0003	99,9907		
58	0,0004	99,9911		
59	0,0004	99,9916		
60	0,0002	99,9918		
61	0,0004	99,9922		
62	0,0004	99,9927		
63	0,0003	99,9930		
64	0,0003	99,9933		
65	0,0009	99,9942		
66	0,0005	99,9947		
67	0,0008	99,9955		
68	0,0004	99,9959		
69	0,0005	99,9965		
70	0,0005	99,9970		

71	0,0007	99,9977		
72	0,0005	99,9982		
73	0,0004	99,9987		
74	0,0009	99,9996		
75	0,0004	100		

	Turc		**Hindi**		**Japonais**	
n	**% flux observé = n**	**% flux observé ≤n**	**% flux observé = n**	**% flux observé ≤n**	**% flux observé = n**	**% flux observé ≤n**
0	0	0	0	0	0	0
1	28,4217	28,4217	19,4423	19,4423	20,3459	20,3459
2	31,0996	59,5213	25,9250	45,3672	31,3401	51,6859
3	21,5977	81,1190	24,5990	69,9663	24,4743	76,1602
4	11,1476	92,2666	16,6558	86,6221	13,3865	89,5467
5	4,8467	97,1133	8,2678	94,8899	5,9166	95,4633
6	1,8493	98,9627	3,3749	98,2648	2,4202	97,8835
7	0,6930	99,6557	1,1841	99,4489	1,0140	98,8975
8	0,2316	99,8874	0,3821	99,8311	0,4696	99,3671
9	0,0706	99,9580	0,1199	99,9509	0,2367	99,6038
10	0,0325	99,9905	0,0331	99,9840	0,1328	99,7366
11	0,0083	99,9987	0,0090	99,9931	0,0782	99,8147
12	0,0013	100	0,0045	99,9976	0,0462	99,8609
13			0,0021	99,9997	0,0321	99,8930
14			0,0003	100	0,0231	99,9161
15					0,0167	99,9328
16					0,0130	99,9458

17					0,0103	99,9561
18					0,0081	99,9642
19					0,0067	99,9709
20					0,0064	99,9773
21					0,0042	99,9815
22					0,0038	99,9852
23					0,0027	99,9879
24					0,0019	99,9898
25					0,0014	99,9912
26					0,0010	99,9922
27					0,0010	99,9932
28					0,0006	99,9938
29					0,0006	99,9944
30					0,0006	99,9950
31					0,0005	99,9954
32					0,0005	99,9959
33					0,0005	99,9964
34					0,0005	99,9969
35					0,0003	99,9972
36					0,0002	99,9974
37					0,0002	99,9976
38					0,0001	99,9977
39					0,0001	99,9978
40					0,0001	99,9978
41					0,0001	99,9979
42					0,0001	99,9980
43					0,0001	99,9981

44					0,0001	99,9982
45					0,0001	99,9982
46					0,0001	99,9983
47					0,0001	99,9984
48					0,0001	99,9985
49					0,0001	99,9986
50					0,0001	99,9986
51					0,0001	99,9987
52					0,0001	99,9988
53					0,0001	99,9989
54					0,0001	99,9990
55					0,0001	99,9990
56					0,0001	99,9991
57					0,0001	99,9992
58					0,0001	99,9993
59					0,0001	99,9994
60					0,0001	99,9994
61					0,0001	99,9995
62					0,0001	99,9996
63					0,0001	99,9997
64					0,0001	99,9998
65					0,0001	99,9998
66					0,0001	99,9999
67					0,0001	100

8.2.4 Flux requis dans SUD

n	% flux requis = n	% flux requis ≤n
0	18,4655	18,4655
1	36,8584	55,3239
2	24,0668	79,3906
3	12,0299	91,4205
4	5,3167	96,7372
5	2,0835	98,8207
6	0,7622	99,5829
7	0,2633	99,8462
8	**0,0906**	**99,9368**
9	0,0323	99,9691
10	0,0136	99,9827
11	0,0063	99,9889
12	0,0034	99,9923
13	0,0022	99,9945
14	0,0014	99,9960
15	0,0009	99,9968
16	0,0007	99,9975
17	0,0005	99,9980
18	0,0004	99,9984
19	0,0003	99,9987
20	0,0002	99,9989

8.2.5 Flux requis dans les cinq langues dans SUD

n	Arabe		Irlandais	
	% flux requis = n	% flux requis ≤n	% flux requis = n	% flux requis ≤n
0	21,5399	21,5399	32,9020	32,9020
1	40,7472	62,2871	42,6465	75,5486
2	23,2620	85,5491	18,3363	93,8849
3	9,6028	95,1519	4,2984	98,1833
4	3,1098	98,2617	1,4512	99,6345
5	1,0151	99,2768	0,2984	99,9329
6	0,3674	99,6442	0,0551	99,9880
7	0,1650	99,8092	0,0090	99,9970
8	0,0826	99,8918	0,0030	100
9	0,0488	99,9406		
10	0,0254	99,9661		
11	0,0153	99,9814		
12	0,0078	99,9892		
13	0,0064	99,9955		
14	0,0036	99,9991		
15	0,0008	99,9999		
16	0,0001	100		

n	Turc		Hindi		Japonais	
	% flux requis = n	% flux requis ≤n	% flux requis = n	% flux requis ≤n	% flux requis = n	% flux requis ≤n
0	4,6883	4,6883	1,4831	1,4831	0,2840	0,2840
1	32,9388	37,6270	20,3996	21,8827	21,0372	21,3212

2	31,2562	68,8832	28,9701	50,8528	31,6705	52,9917
3	18,9993	87,8825	24,8125	75,6654	24,2172	77,2089
4	8,2298	96,1123	14,7749	90,4403	12,9724	90,1813
5	2,7823	98,8946	6,4688	96,9091	5,6332	95,8145
6	0,8057	99,7003	2,1911	99,1002	2,2763	98,0908
7	0,2266	99,9268	0,6595	99,7597	0,9449	99,0357
8	0,0503	99,9771	0,1810	99,9407	0,4311	99,4668
9	0,0165	99,9936	0,0404	99,9810	0,2087	99,6755
10	0,0057	99,9994	0,0123	99,9934	0,1148	99,7903
11	0,0006	100	0,0051	99,9985	0,0645	99,8548
12			0,0015	100	0,0376	99,8924
13					0,0247	99,9171
14					0,0173	99,9345
15					0,0129	99,9474
16					0,0106	99,9580
17					0,0081	99,9660
18					0,0062	99,9722
19					0,0052	99,9774
20					0,0042	99,9815
21					0,0028	99,9843
22					0,0024	99,9867
23					0,0021	99,9888
24					0,0015	99,9903
25					0,0012	99,9915
26					0,0010	99,9925
27					0,0008	99,9933
28					0,0006	99,9938

29					0,0006	99,9944
30					0,0006	99,9950
31					0,0005	99,9954
32					0,0005	99,9959
33					0,0005	99,9964
34					0,0005	99,9969
35					0,0003	99,9972
36					0,0002	99,9974
37					0,0002	99,9976
38					0,0001	99,9977
39					0,0001	99,9978
40					0,0001	99,9978
41					0,0001	99,9979
42					0,0001	99,9980
43					0,0001	99,9981
44					0,0001	99,9982
45					0,0001	99,9982
46					0,0001	99,9983
47					0,0001	99,9984
48					0,0001	99,9985
49					0,0001	99,9986
50					0,0001	99,9986
51					0,0001	99,9987
52					0,0001	99,9988
53					0,0001	99,9989
54					0,0001	99,9990
55					0,0001	99,9990

56					0,0001	99,9991
57					0,0001	99,9992
58					0,0001	99,9993
59					0,0001	99,9994
60					0,0001	99,9994
61					0,0001	99,9995
62					0,0001	99,9996
63					0,0001	99,9997
64					0,0001	99,9998
65					0,0001	99,9998
66					0,0001	99,9999
67					0,0001	100