Entre arbre de dépendance et ordre linéaire, les deux processus de transformation : linéarisation, puis reconstruction de l'arbre

Jacques Vergne*

Résumé - Abstract

Dans cet article, nous nous situons en linguistique informatique et nous présentons des hypothèses sur les deux processus qui permettent à une personne de coder un arbre en une phrase, et à une autre personne de décoder cette phrase en reconstruisant l'arbre, et nous y associons les deux formes structurelles que peut prendre une phrase : son ordre linéaire et son arbre de dépendance. Nous faisons l'hypothèse que cette présentation simultanée, cette appréhension des structures sous les contraintes des processus, permettent une meilleure théorisation de l'ensemble processus - structures.

In this paper, dealing with computational linguistics, we present hypotheses on the two processes which make a human being able to code a tree into a linear sentence, and another human being able to decode this sentence while rebuilding the tree, and we also present the two structural forms a sentence can be represented in : its linear order and its dependency tree. We make the hypothesis that this simultaneous presentation of structures through processes allows a better theorisation of processes and structures.

* GREYC (UPRESA 6072), Université de Caen, F-14032 Caen cedex
courrier électronique : Jacques.Vergne@info.unicaen.fr

0. Introduction

Notre domaine de recherches se situe en linguistique informatique[1], plus précisément en syntaxe, en tant qu'**étude des formes indépendamment du sens**. Ces recherches en syntaxe sont effectuées à l'aide de l'ordinateur (voir ci-dessous en 3.2.3.5.). Dans notre démarche, nous pensons qu'il est plus intéressant et plus faisable d'expliciter les *processus* qui produisent ou interprètent les structures, que de poursuivre sans fin l'énumération exhaustive des *structures*, y compris exprimées sous la forme condensée des grammaires formelles. En outre, cette démarche est favorisée par le fait que l'ordinateur, notre principal outil de recherche, est plus un outil d'explicitation et d'exécution de processus (les programmes, les algorithmes), que de reconnaissance de structures supposées connues au moment de la conception des algorithmes. Ceci est une intrusion explicite de la chronologie, précédemment mise à l'écart quoique présente au cœur même de l'outil de recherche, l'ordinateur. L'évolution proposée consiste donc à expliciter des processus qui observent, extraient, produisent les structures des phrases analysées, sans attendu préalable sur ces structures. En généralisant, on peut dire qu'un objet statique devient plus intelligible si l'on inclut dans l'étude les processus qui l'ont produit et les processus qui l'exploitent[2]. Bien entendu, les deux processus en jeu ici sont ceux de la production et de la réception dans la situation de communication de deux êtres humains. Ajoutons qu'une telle étude de l'objet, qui inclut les processus *ante* et *post*, fait émerger dans l'objet des traces des processus; plus précisément on trouve dans l'objet observé des marques de *contraintes* qui se sont appliquées dans le processus de production, et des marques de contraintes qui s'appliqueront dans le processus de réception, pour le rendre possible et le faciliter.

Cet article tente ainsi de replacer la phrase linéaire entre le processus qui l'a produite et celui qui va l'interpréter, en étudiant comme un tout la phrase linéaire (objet observable) avec ces deux processus (hypothèses).

Dans la section 1, nous présentons les deux processus successifs : production puis réception de la phrase (les processus sur la phrase), ou bien linéarisation puis reconstruction de l'arbre (ces mêmes processus, mais sur l'arbre). La section 2 concerne les structures : présentation des segments reliés (les syntagmes réduits), présentation de l'arbre de dépendance à transmettre, et de la phrase linéaire transmise, c'est-à-dire l'arbre de dépendance linéarisé. La section 3 présente et détaille les processus **:** le processus de production (l'opération de linéarisation optimisée de l'arbre) et le processus de réception (la reconstruction de l'arbre par mise en relation des syntagmes). Dans les deux processus, les contraintes mémorielles jouent un rôle central.

---

[1] Entendons par *linguistique informatique* un domaine de recherches en linguistique où l'informatique joue un rôle d'instrument de recherche, à distinguer de l'*informatique linguistique*, domaine de recherches en informatique où un matériau linguistique est l'objet traité, même si ces deux domaines sont liés concrètement.

[2] Par exemple, la composition du lait n'est intelligible que si l'on a compris qu'un mammifère produit une nourriture pour sa progéniture.

1. Les deux processus successifs : linéarisation puis reconstruction de l'arbre

Lucien Tesnière définit les deux "ordres" structural et linéaire de la manière suivante dans (Tesnière 59), page 16, § 1 :

> *1.- L'ordre structural des mots est celui selon lequel s'établissent les connexions.*

et page 18, § 8 :

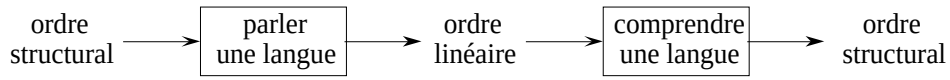
> *8.- Nous appellerons ordre linéaire celui d'après lequel les mots viennent se ranger sur la chaîne parlée. L'ordre linéaire est, comme la chaîne parlée, à une dimension.*

Ensuite, page 19, § 4, il présente ainsi les processus :

> *[...] nous pouvons dire que [...]* ***parler*** *une langue, c'est en transformer l'ordre structural en ordre linéaire, et inversement que* ***comprendre*** *une langue, c'est en transformer l'ordre linéaire en ordre structural.*

Remarquons qu'il les présente comme des processus réciproques, et non pas comme des processus successifs dans la situation de communication, ce qui donne :

Figure 1 : Les deux processus selon Tesnière, présentés comme successifs

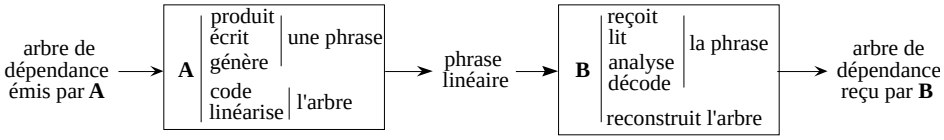


Pour ces mêmes processus, (Grunig 93), page 18, propose les termes d'"encodage" et de "décodage" :

> *Par encodage j'entends évidemment la projection de structures abstraites arborescentes (en l'occurrence dépendancielles) dans la linéarité des chaînes énoncées. Par décodage, inversement, j'entends un mode de récupération de l'arborescence qui est rendu possible par les chaînes obtenues.*

Voici maintenant comment nous concevons les deux processus, placés en situation de transmission d'information : une personne **A** transmet un arbre de dépendance à une personne **B**, en le codant temporairement sous forme de phrase linéaire. Les processus sont désignés en tant qu'actions exécutées par A puis par B, et s'appliquant soit à l'arbre soit à la phrase :

Figure 2 : Les deux processus successifs, placés en situation de transmission d'un arbre de dépendance



Les deux processus sont successifs, réciproques du point de vue fonctionnel, mais différents du point de vue organique, et exécutés par deux personnes différentes. L'arbre de dépendance est une représentation abstraite hypothétique, alors que la phrase linéaire est le matériau observable.

Au cours de ces deux processus successifs, l'arbre de dépendance est ainsi l'objet, l'information à transmettre entre les deux personnes qui communiquent. Cet arbre de dépendance est temporairement linéarisé (et codé, compressé) dans la phrase linéaire.

Ce sont évidemment des êtres humains qui exécutent ces processus, mais il est possible de les modéliser sur ordinateur, et ceci avec deux objectifs possibles : soit un objectif opératoire de génération ou d'analyse de phrase dans le cadre d'une application informatique plus large (traduction automatique, indexation automatique, correction orthographique), soit un objectif de recherche fondamentale sur la syntaxe, qui est pour nous l'objectif prioritaire. Cette modélisation est une réduction, une schématisation, une analogie des processus, à usage exploratoire et expérimental.

Les processus et les structures sont façonnés par de nombreuses contraintes :

- contrainte des propriétés géométriques des structures, surtout du fait que la phase linéaire est unidimensionnelle;
- contrainte du fait que l'information contenue dans l'arbre n'est pas perdue au cours de la linéarisation : cette information est codée différemment dans la phrase linéaire;
- contrainte de la chronologie des processus et des déductions;
- contrainte de la minimisation de l'effort de mémoire (voir ci-dessous en 3.1.), qui implique la minimisation des distances entre nœuds dans l'ordre linéaire[3].

Les deux structures sont difficilement théorisables statiquement, en dehors des deux processus, comme le montrent les tentatives des linguistes et des chercheurs en traitement automatique des langues (TAL). Les linguistes ont surtout tenté de modéliser arbre et phrase linéaire de manière statique : la

[3] Cette hypothèse du moindre effort est déjà ancienne : on la trouve déjà chez (Zipf 1949), et plus récemment chez (Grunig 1993), page 18, au sujet du moindre effort de mémoire : "*Il apparaîtra que certains types d'encodages (et décodages associés) sont, du seul point de vue des charges mémorielles que je retiens ici (...), clairement plus économiques que d'autres*".

génération chez Chomsky reste le processus de déduction de la démarche hypothético-déductive, (Mel'cuk 88) dit page 129 sa difficulté à définir sa "*syntactic dependency*", et la connexion chez Tesnière est définie comme un processus de perception et de calcul (voir ci-dessous en 3.2.2.) mais le concept d'arbre de dépendance (le stemma) est resté statique, et la correspondance entre ordre linéaire et ordre structural, présentée sous la forme des deux processus, n'est pas développée, mais seulement esquissée par le terme de "image projetée", uniquement dans le paragraphe suivant, page 20, § 10 :

> *[...] syntaxiquement, la vraie phrase, c'est la* ***phrase structurale*** *dont la phrase linéaire n'est que l'image projetée tant bien que mal, et avec tous les inconvénients d'aplatissement que comporte cette projection sur la chaîne parlée.*

Quant à l'algorithme de l'analyse syntaxique automatique classique, c'est un processus combinatoire d'essais de tous les choix possibles jusqu'à la *reconnaissance de structures* explicitement attendues, processus analogue à la compilation, qui ne modélise pas un *processus de mise en relation* (alors que c'est le principe même du processus de reconstruction de l'arbre présenté ci-dessous en 3.2.).

Notre démarche est de faire des hypothèses sur l'ensemble des deux structures et des deux processus, des hypothèses explicites sur les processus, qui conduisent à définir implicitement les structures à partir des processus. Dans le travail présenté ici, nous proposons :

- pour le processus de linéarisation de l'arbre (en 3.1.), des hypothèses sur le parcours d'arbre issues de l'informatique théorique, confrontées à une étude sur corpus ;
- pour le processus de reconstruction de l'arbre (en 3.2.), des hypothèses sur un processus de mise en relation fondé sur la mémoire, modélisées et généralisées dans le cadre d'un analyseur syntaxique, puis confrontées à l'analyse automatique d'un corpus.

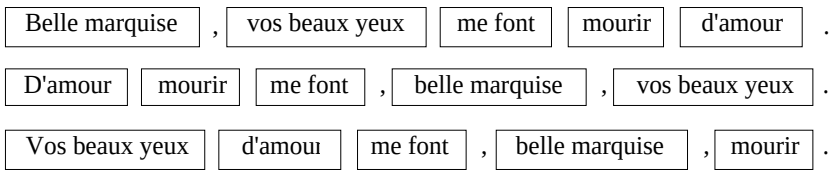
2. Structures

Cette section présente d'abord (en 2.1.) le segment (le syntagme réduit) qui est le constituant de base des deux structures : l'arbre de dépendance à transmettre (2.2.), et la phrase linéaire transmise, c'est-à-dire l'arbre de dépendance linéarisé (2.3.).

2.1. Les segments reliés = les Syntagmes Réduits (SR)

Observons comment, dans Le Bourgeois Gentilhomme (figure 3), Molière fait permuter des groupes de mots dans la phrase, mais laisse inchangé l'ordre des mots dans chaque groupe :

Figure 3 : Définition implicite d'un segment permutable par Molière



Le segment de phrase choisi pour être le segment relié par les relations de dépendance est non pas le mot, dont la pratique est conventionnelle et instable entre des langues différentes, mais ce groupe de mots que Molière fait permuter : le syntagme réduit, ou syntagme simple, ou syntagme noyau, ou "core phrase", ou "chunk" dans la littérature en anglais (voir Abney 1996), ou syntagme non récursif, ou syntagme sans ses syntagmes compléments, par opposition au syntagme récursif des grammaires génératives, qui inclut ses syntagmes compléments.

Dans la suite de cet article, nous appellerons ce segment SR, comme Syntagme Réduit. Il sera délimité par un rectangle dans toutes les figures, ce qui donnera au lecteur des exemples par la pratique. Sans le définir par l'inventaire exhaustif de ses structures[4], donnons-en une description générale : un SR est constitué d'un élément central, le plus souvent un nom (ou un pronom tonique) ou un verbe (conjugué, infinitif, participe présent ou passé), entouré éventuellement de ses éléments périphériques :

- dans le SR nominal : conjonction de coordination et/ou de subordination, préposition, déterminant, adjectif épithète antéposé ou postposé, adverbe antéposé à l'adjectif épithète; exemples : de *avec appétit* à *mais avec une très belle raquette blanche* ;
- dans le SR verbal : conjonction de coordination et/ou de subordination, préposition, tous les pronoms atones (sujet, objet et autres) antéposés ou postposés, négations, auxiliaire, adverbe le plus souvent postposé, adjectif attribut avec la copule être, adverbe antéposé à l'adjectif attribut; exemples : de *ne voit* à *qu'il ne le leur a sûrement pas donné hier*.

L'élision est interne au SR (le mot élidé et le mot suivant appartiennent au même SR). Une ponctuation ne peut couper un SR. Le SR nominal est marqué par un genre et un nombre homogènes sur tous ses composants variables en flexion. Ce segment montre sa stabilité sur des langues variées

[4] La question de la *définition d'un segment linguistique* est délicate. Nous proposons de l'aborder de la même manière que la question des structures de phrase : ne cherchons pas à donner une définition exhaustive d'un segment par sa **structure** en termes d'autres segments (par exemple sous la forme d'une grammaire formelle), mais cherchons à définir et expliciter exhaustivement un **processus** d'analyse qui délimite un tel segment dans le flux textuel entrant (par des connaissances sur son début et sa fin), qui l'identifie et produit sa structure; autrement dit, ne donnons pas de définition directe a priori, mais une définition indirecte par un processus d'analyse explicité, donc modélisable et automatisable.

comme on peut le voir dans les travaux de Hervé Déjean (cf. Déjean 1998a, en 3.1., 1998b).

En se référant aux définitions possibles du "groupe verbal" par (Le Goffic 1993), page 29, le SR ne correspond ni à sa "définition large" (le verbe avec tous ses compléments = le prédicat complet = le syntagme verbal récursif), ni à sa "définition étroite" (le verbe avec son auxiliaire éventuel, en excluant tout pronom atone), mais se situe entre les deux[5].

Le SR permet de définir une hiérarchie de segments à trois niveaux : mots, SR, phrases, hiérarchie où le segment d'un niveau est constitué de segments du niveau inférieur : le type d'un tout est différent du type de ses parties, contrairement au syntagme récursif, constitué de mots et de syntagmes récursifs. Les mots dans un SR, et les SR dans une phrase ont des comportements très différents : les mots d'un SR sont dans un ordre très contraint autour d'un nom ou d'un verbe, mais les SR dans une phrase sont dans un ordre soumis à des contraintes plus relâchées.

Notons que, dans notre démarche, constituance[6] et dépendance ne sont ni déclarées équivalentes (Gaifman 1965), ni opposées, mais utilisées ensemble à deux niveaux différents : constituance à l'intérieur des SR, dépendance entre SR.

2.2. L'arbre de dépendance à transmettre

Dans l'arbre de dépendance à transmettre, théoriquement identique à l'arbre reçu, les segments reliés (i.e. les nœuds de l'arbre) sont des SR (et non pas des mots comme pour Tesnière), et ils sont reliés par des relations de dépendance. Le nœud-racine conventionnel est le SR verbal de la proposition principale.

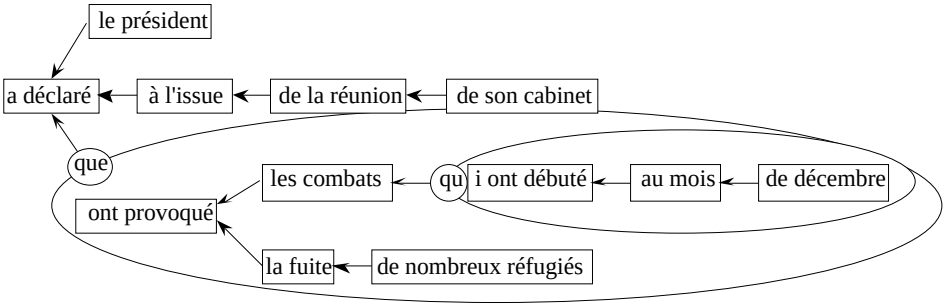
En voici un exemple, dans la figure 4 ci-dessous, où l'arbre est dessiné horizontalement, pour maintenir l'écriture horizontale des SR et des suites de SR dépendants, et dans lequel le SR-racine est conventionnellement placé à gauche. Remarquons que nous avons toute liberté de dessin de l'arbre, car nous nous libérons de la contrainte de projectivité (la projectivité est la propriété d'être projetable pour un arbre de dépendance) en concevant la linéarisation différemment (par parcours de l'arbre, voir ci-dessous en 3.1.1.); cet abandon de la contrainte de projectivité est en quelque sorte un retour au stemma de Tesnière, qui représente l'ordre structural dissocié de l'ordre linéaire.

---

[5] Remarque : notre définition "médiane" du SR correspond à une définition possible du groupe accentuel (qui contient un seul accent primaire, et éventuellement des accents secondaires); les phénomènes d'élision, liaison obligatoire et enchaînement sont internes au groupe accentuel ainsi défini et permettent des tests en cas d'hésitation sur ses limites.

[6] Nota bene : pour ce néologisme, on trouve les deux orthographes *constituance* et *constituence* dans la littérature. L'orthographe *constituance* est plus fréquente. Nous l'avons préférée en tant que dérivée de *constituant*, comme *dépendance* est dérivé de *dépendant*, alors que *constituence* est un anglicisme à partir de *constituency*, dérivé de *constituent*, comme *dependency* est dérivé de *dependent*.

Figure 4 : L'arbre de dépendance à transmettre



Nous proposons une représentation particulière de la proposition subordonnée : le régissant de sa hiérarchie interne (appelé souvent sa "tête") est son SR verbal conjugué; mais, vue de son extérieur, la proposition subordonnée constitue un tout (entouré d'un ovale), relié à son régissant par l'intermédiaire de la conjonction de subordination (entourée d'un cercle) qui sera linéarisée au début de la subordonnée. On dissocie ainsi deux questions habituellement liées : d'une part la subordonnée inclut le régissant interne de sa hiérarchie interne, et d'autre part c'est la subordonnée entière (et non pas un de ses éléments) qui dépend de son régissant externe (comme le SR entier dépend de son régissant externe)[7].

En ce qui concerne la notion de dépendance, nous la considérons de manière classique comme une notion générique de la complémentation et de la subordination.

2.3. La phrase linéaire transmise = l'arbre de dépendance linéarisé

La phrase linéaire transmise est le résultat du processus de linéarisation de l'arbre de dépendance (décrit ci-dessous en 3.1). Les segments et les dépendances sont les mêmes que dans l'arbre à transmettre.

Pour pouvoir mesurer les longueurs des dépendances, et mesurer et comparer des longueurs de constituants, nous avons défini une métrique sur la phrase linéaire de la manière suivante :

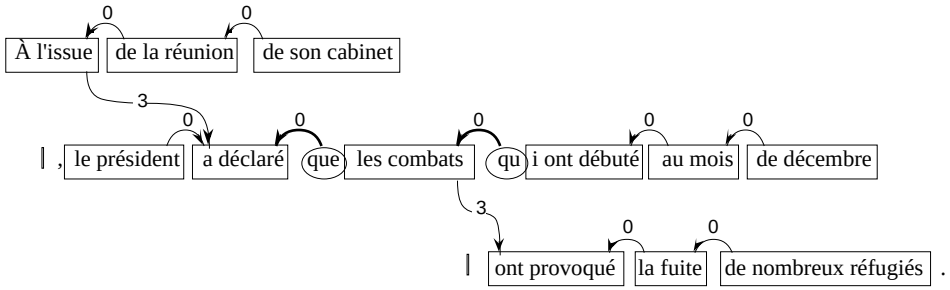
- l'unité de la métrique est le Syntagme Réduit (SR), délimité par un rectangle dans les figures ;
- la longueur de la dépendance entre 2 SR est définie comme étant le nombre de SR qui les séparent, ce qui implique que la longueur de la dépendance entre 2 SR dépendants contigus est nulle ;

[7] Voir une représentation analogue : les "arbres à bulles" de (Kahane 97). Voir aussi la translation du second degré de Tesnière : "La translation I>>O" (qui translate un verbe en substantif) dans (Tesnière 1959), page 546, chapitre 241, translation dont le translatif est la conjonction de subordination *que*.

- la longueur d'un groupe de SR contigus est le nombre de SR qu'il contient ;
- la longueur de la dépendance entre une proposition subordonnée et son SR régissant est définie comme étant le nombre de SR qui séparent ce régissant du début (ou de la fin, s'il est situé après) de la proposition subordonnée, ce qui implique que cette longueur est nulle si ces deux constituants reliés sont contigus (la contiguïté entre une proposition subordonnée et son SR régissant est donc définie comme la contiguïté entre 2 SR); en étant définie ainsi, la longueur de la dépendance entre une proposition subordonnée et son SR régissant reste indépendante de la linéarisation interne de la subordonnée (cette dépendance particulière est marquée par une flèche plus épaisse dans les figures des arbres linéarisés).

En marquant chaque dépendance par une flèche accompagnée de la longueur de la dépendance, on obtient alors l'arbre de dépendance linéarisé de la figure 5.

Figure 5 : Phrase linéaire transmise = arbre de dépendance linéarisé[8]



3. Processus

Nous allons présenter dans cette section nos hypothèses sur les deux processus de production et de réception, dans l'ordre chronologique de la transmission de l'arbre de dépendance émis, linéarisé dans la phrase linéaire (3.1.), puis ensuite reconstruit (3.2.).

[8] Remarque sur la figure 5 : cette figure est disposée en trois lignes, et chaque ligne est une suite de SR dépendants contigus (reliés par une dépendance de longueur nulle), ce qui constitue une définition possible du groupe prosodique, c'est-à-dire une suite de groupes accentuels dits d'un seul tenant, sans pause; la frontière entre deux groupes prosodiques est placée entre deux SR contigus non directement reliés (frontière marquée par le signe | dans la figure 5, et dans la suite de cet article), et est marquée à l'oral par une pause de durée d'autant plus longue que la dépendance est longue; ceci établit un lien précis entre la structure de l'arbre de dépendance linéarisé et cette définition du groupe prosodique.

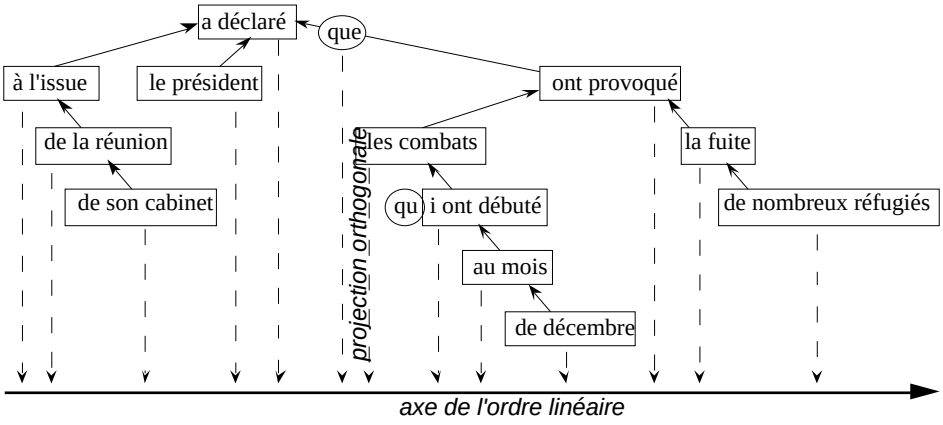
3.1. Le processus de production : arbre → phrase linéaire = opération de linéarisation optimisée

Nous allons faire une hypothèse sur le processus par lequel, à la production, la phrase linéaire est produite à partir de l'arbre de dépendance : comment un régissant peut être linéarisé par rapport à ses dépendants, comment les dépendants sont linéarisés, et enfin quels sont les critères de calcul de la linéarisation ? Nous terminerons cette section par une étude de la linéarisation sujet - verbe dans les relatives en *que*, étude théorique fondée sur les concepts exposés dans cette sous-section, puis corroborée sur corpus.

3.1.1. Linéariser l'arbre = le parcourir en relevant les nœuds

La linéarisation classique de l'arbre de dépendance consiste en sa projection géométrique orthogonale sur l'axe de l'ordre linéaire, qui a été proposée par (Hays 1964), puis adoptée dans la communauté des Traitements Automatiques des Langues (TAL). Pour rendre projetable l'arbre de la figure 4, il faudrait le dessiner comme dans la figure 6 ci-dessous, laborieusement, car sous deux contraintes difficilement compatibles (et quelquefois incompatibles) : placer un dépendant à la fois au dessous de son régissant et après le syntagme contigu précédent, tout en maintenant la projectivité de l'arbre.

Figure 6 : Arbre de dépendance projetable à la manière de Hays



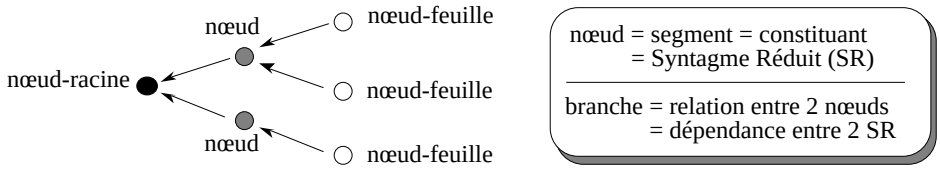
La linéarisation par projection a au moins les trois inconvénients suivants : 1) l'ordre linéaire dépend de la manière dont l'arbre est dessiné; 2) elle a donné naissance au concept de projectivité, qui était un intermédiaire nécessaire pour montrer l'équivalence formelle "*entre la génération par un système de constituance et la génération par un système de dépendance*" possédant un vocabulaire non terminal et reposant sur la projectivité - voir (Portine 1992), page 123, et (Gaifman 1965); et 3) la non-projectivité de phrases extraites de

corpus est avérée (par exemple en cas d'incise). L'arbre projetable, contrairement au stemma, a ainsi l'ambition de coder simultanément dans un même dessin à la fois l'ordre structural et l'ordre linéaire (comme l'arbre syntagmatique d'ailleurs).

Nous proposons que l'ordre linéaire soit non pas lisible directement sur l'arbre projetable, mais qu'il soit le résultat d'un processus sur la structure de l'arbre à transmettre. Nous abandonnons alors simplement la projection pour le parcours d'arbre, qui ne dépend pas du dessin de l'arbre, mais uniquement de sa structure.

Produire une phrase, c'est placer les segments reliés sur l'axe de la phrase (espace à une dimension), c'est-à-dire énumérer les nœuds de l'arbre de dépendance dans un certain ordre. L'informatique théorique apporte des concepts clairs et adéquats au sujet des arbres et des ordres possibles d'énumération des nœuds d'un arbre : le processus de parcours d'arbre avec relevé des nœuds. La figure 7 rappelle les éléments d'un arbre, et la définition des nœuds et branches de l'arbre de dépendance :

Figure 7 : Nœuds et branches de l'arbre de dépendance



Parcourir un arbre, c'est passer par tous ses nœuds en suivant un certain chemin, en partant de la racine, et en revenant à la racine; relever les nœuds d'un arbre, c'est parcourir l'arbre en "ramassant" au passage chaque nœud une seule fois, ce qui revient à énumérer les nœuds de l'arbre dans un certain ordre.

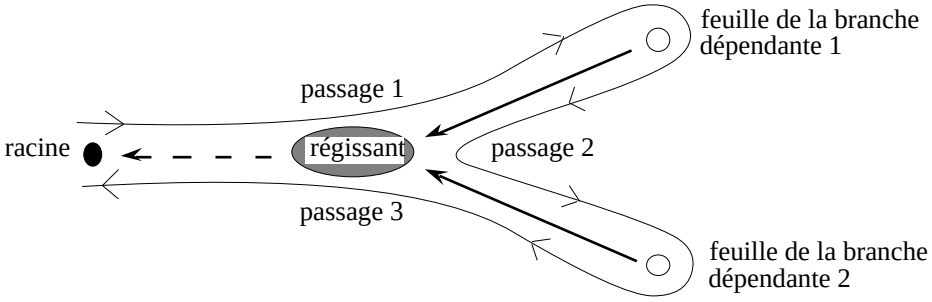
Nous pouvons alors considérer que linéariser un arbre revient à le parcourir en relevant chaque nœud une fois.

On observe que ce relevé est le plus souvent celui qui place les nœuds reliés les plus proches possible (contigus si possible) dans l'ordre linéaire, et nous faisons l'hypothèse que c'est pour **minimiser** les efforts de mémoire à la production et à la réception.

Le parcours d'arbre qui minimise les distances entre nœuds s'appelle le parcours en profondeur d'abord : aller de la racine vers les feuilles, et retour des feuilles vers la racine, en suivant toutes les branches. Pour un nœud à deux branches, une première branche est parcourue en entier avant de parcourir la deuxième branche en entier.

Dans un tel parcours d'arbre, on passe trois fois sur un nœud-régissant qui a deux branches dépendantes (voir figure 8) :

Figure 8 : Parcours d'arbre en profondeur d'abord



aller : racine → nœud-régissant → feuille de la branche dépendante 1,
retour : feuille de la branche dépendante 1 → nœud-régissant,
aller : nœud-régissant → feuille de la branche dépendante 2,
retour : feuille de la branche dépendante 2 → nœud-régissant → racine.

Le type de **relevé** du nœud régissant spécifie à quel passage il est relevé :

- au passage 1, le régissant est relevé **avant** ses dépendants (relevé dit **préfixé**)
- au passage 2, le régissant est relevé **entre** ses dépendants (relevé dit **infixé**)
- au passage 3, le régissant est relevé **après** ses dépendants (relevé dit **postfixé**).

Dans un même parcours d'arbre, chaque nœud a son propre type de relevé. En français, le relevé normal est le relevé préfixé : linéarisation régissant puis dépendants, sauf le verbe qui est infixé entre son sujet et ses autres compléments.

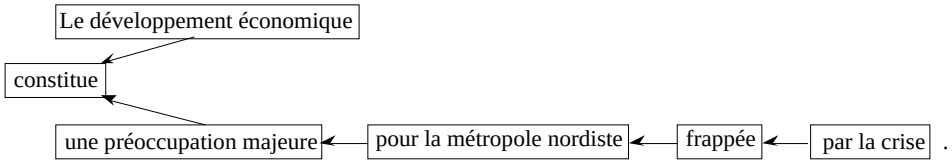
Ces concepts vont nous permettre de catégoriser les différents ordres linéaires possibles entre un nœud régissant et ses nœuds dépendants (voir ci-dessous en 3.1.2.).

3.1.2. Linéarisation d'un nœud régissant par rapport à ses dépendants

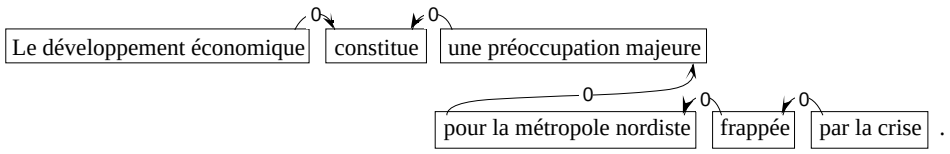
Quand, dans une phrase, un nœud A est relié uniquement à un nœud B, le nœud B peut être placé avant ou après A : A - B ou B - A et les deux nœuds reliés sont contigus; quand un nœud A est relié uniquement à deux nœuds B et C, ces deux nœuds peuvent être placés autour de A : B - A - C ou C - A - B et les nœuds reliés sont toujours contigus deux à deux. Dans la figure 9, voici un exemple de phrase où les nœuds sont reliés à un ou deux autres nœuds, et où toute relation peut ainsi être marquée par une contiguïté :

Figure 9 : Phrase où tous les nœuds reliés sont contigus

arbre de dépendance à transmettre :



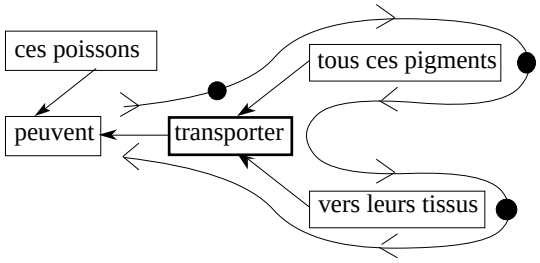
arbre de dépendance linéarisé = phrase linéaire :



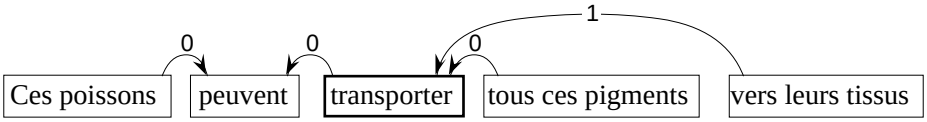
Mais quand un nœud A est relié à trois autres nœuds B, C et D, il n'y a que deux contiguïtés disponibles autour de A, et une des trois relations à A ne peut se faire en contiguïté avec A : c'est une contrainte de l'espace à une dimension, ou bien un des "*inconvénients d'aplatissement*" dont parle Tesnière. En particulier, quand un nœud régissant (qui a lui-même son régissant) a deux nœuds dépendants, ils ne peuvent le suivre tous les deux en contiguïté, alors trois relevés du nœud régissant sont possibles. Dans les exemples des figures 10, 11 et 12), nous allons présenter un même arbre de dépendance, linéarisé de 3 manières différentes :

Figure 10 : Nœud régissant relevé au passage 1, donc **préfixé** avant ses nœuds dépendants

parcours de l'arbre de dépendance à transmettre (● = relevé du nœud):



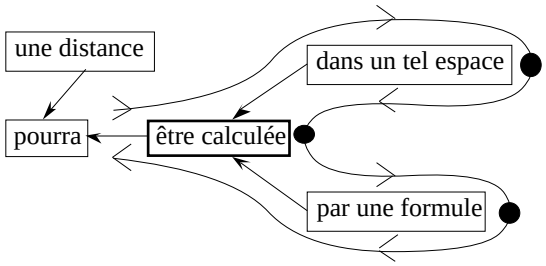
arbre de dépendance linéarisé = phrase linéaire :



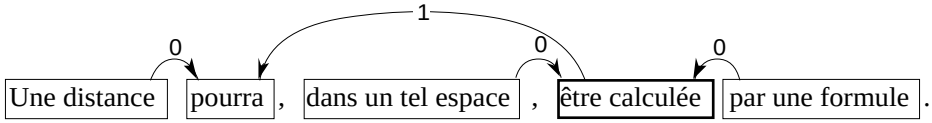
L'actant objet du verbe lui est contigu (à distance nulle), et le complément circonstant du verbe est à distance 1, après l'objet.

Figure 11 : Nœud régissant relevé au passage 2, donc **infixé** entre ses dépendants

parcours de l'arbre de dépendance à transmettre (● = relevé du nœud):



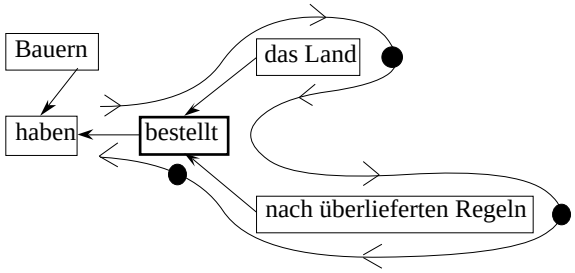
arbre de dépendance linéarisé = phrase linéaire :



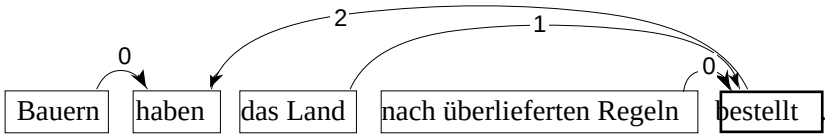
Les deux compléments circonstants de *être calculée* lui sont contigus, et son régissant *pourra* est à distance 1.

Figure 12 : Nœud régissant relevé au passage 3, donc **postfixé** après ses nœuds dépendants

parcours de l'arbre de dépendance à transmettre (● = relevé du nœud):



arbre de dépendance linéarisé = phrase linéaire :



Le cas du participe passé postfixé est classique en allemand : *bestellt* est contigu à son complément circonstant, à distance 1 de son objet, et à distance 2 de son régissant *haben*.

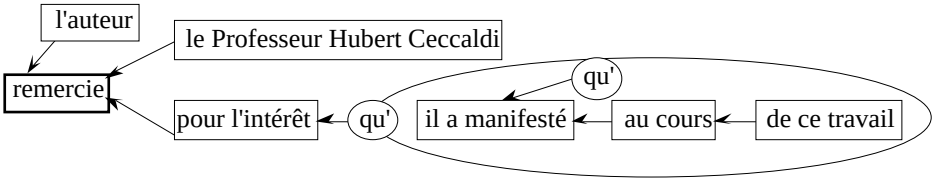
3.1.3. Linéarisation des nœuds dépendants

Demandons-nous maintenant quel est l'ordre des nœuds dépendants après le nœud régissant préfixé. Prenons comme exemple le cas où un verbe (nœud régissant préfixé) a un actant objet direct et un complément circonstant indirect (deux nœuds dépendants). Les deux ordres possibles sont soit verbe puis branche objet puis branche complément circonstant, soit verbe puis branche complément circonstant puis branche objet, chaque branche étant parcourue en entier (figures 13 et 14).

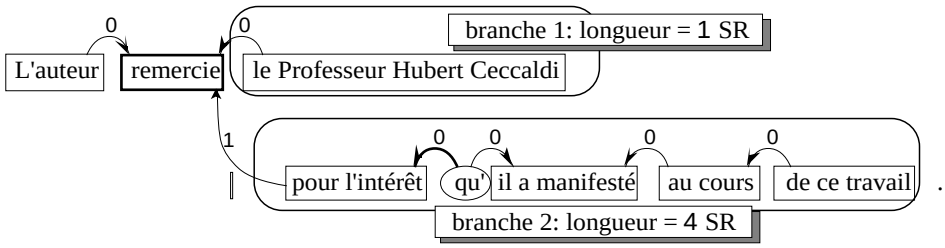
Dans l'arbre de dépendance à transmettre de la figure 13, le pronom relatif *que* est présent deux fois : en tant que conjonction de subordination, et en tant que pronom objet.

Figure 13 : Linéarisation : verbe, puis branche objet, puis branche circonstant

arbre de dépendance à transmettre :



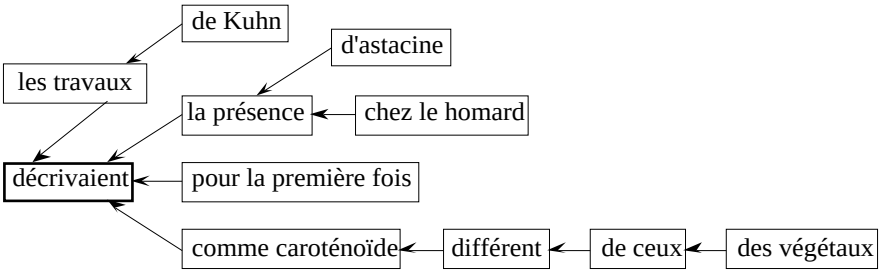
arbre de dépendance linéarisé = phrase linéaire :



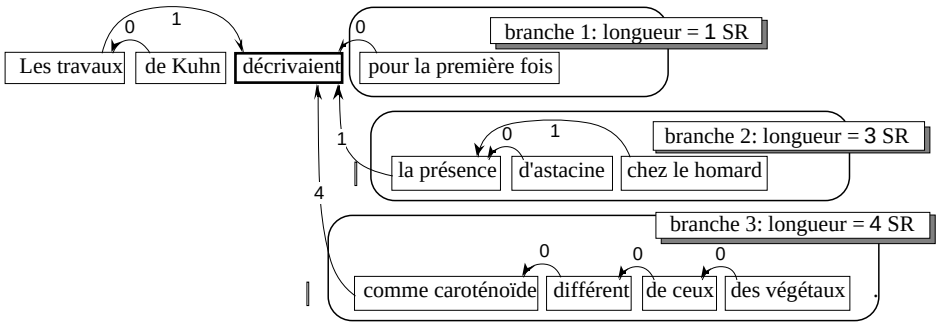
Dans la phrase linéaire de la figure 13, les longueurs des deux branches dépendantes du verbe *remercie* sont mises en évidence: successivement longueurs 1 puis 4.

Figure 14 : Linéarisation : verbe, puis branche circonstant, puis branche objet

arbre de dépendance à transmettre :



arbre de dépendance linéarisé = phrase linéaire :



Dans la phrase linéaire de la figure 14, les longueurs des trois branches dépendantes du verbe *décrivaient* sont mises en évidence: successivement longueurs 1 puis 3 puis 4.

Dans les deux cas, la branche la plus courte est linéarisée la première : des deux ordres linéaires possibles entre deux branches, c'est celui qui minimise la somme des distances entre nœuds reliés. Cette minimisation des distances permet une minimisation de l'effort de mémoire à l'émission (et à la réception) et permettra le calcul des relations à la réception.

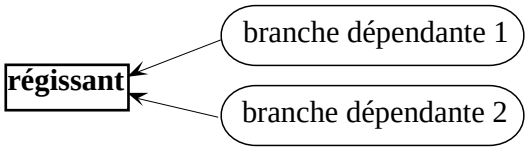
Nous appellerons ce processus "linéarisation optimisée de l'arbre de dépendance".

3.1.4. Calcul du choix de la linéarisation optimisée

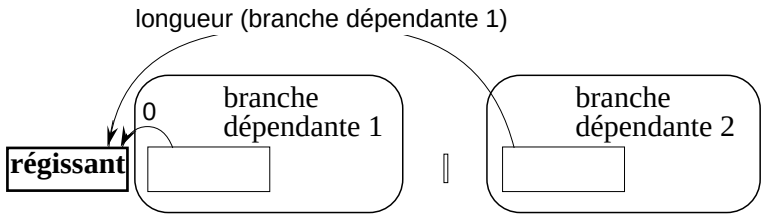
Nous allons maintenant démontrer cette propriété dans le cas normal en français du nœud régissant préfixé, en prenant le cas général abstrait d'un nœud régissant suivi de ses deux nœuds dépendants (un SR régissant préfixé à ses deux dépendants). Dans la figure 15, la branche 1 est conventionnellement linéarisée la première.

Figure 15 : Linéarisation d'un nœud régissant à deux branches dépendantes

arbre de dépendance à transmettre :



linéarisation : régissant puis dépendants, par relevé préfixé



La longueur de la dépendance du premier nœud de la branche dépendante 1 est 0 car cette branche est produite la première, contiguë à son régissant. La longueur de la dépendance du premier nœud de la branche dépendante 2 est longueur (branche dépendante 1) car cette branche est produite dès que la branche 1 se termine.

L'hypothèse du **moindre effort de mémoire** nous conduit à une définition géométrique du critère d'optimisation de la linéarisation :

la linéarisation optimisée est celle qui, parmi toutes les linéarisations possibles, minimise la somme des longueurs des dépendances.

Pour l'ordre linéaire branche 1 suivie de la branche 2 (cas de la figure 15), la somme des longueurs des dépendances est : longueur (branche dépendante 1).

Pour l'ordre linéaire branche 2 puis branche 1, la somme des longueurs des dépendances est : longueur (branche dépendante 2).

La linéarisation optimisée est donc l'ordre linéaire branche 1 puis branche 2 si :

longueur (branche dépendante **1**) ≤ longueur (branche dépendante **2**)

La linéarisation optimisée place en premier la branche la plus courte (pour un nœud régissant suivi de ses deux nœuds dépendants).

En généralisant à un nombre quelconque de branches, la linéarisation optimisée est donc l'ordre linéaire qui place les branches par ordre croissant de longueur : c'est le cas de la figure 14 où l'on a l'ordre linéaire des trois branches de longueurs 1, 3 et 4.

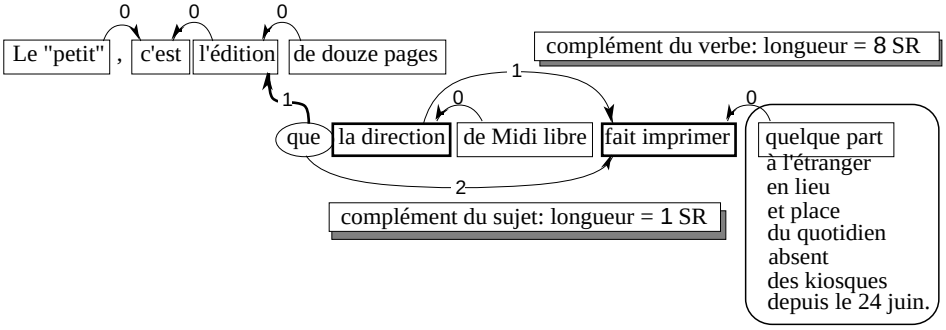
Qu'avons nous démontré ? Nous avons seulement démontré que la minimisation de la somme des distances entre éléments reliés implique un certain ordre des constituants, fonction des longueurs des constituants. Quant à savoir d'où vient la minimisation de la somme des distances entre éléments reliés, nous pouvons seulement émettre l'hypothèse du moindre effort de mémoire (cf. notes 3 et 12) et corroborer (ou falsifier)[9] sur le matériau cette hypothèse par ses conséquences sur l'ordre des constituants (voir ci-dessous l'étude sur corpus en fin de 3.1.5.).

3.1.5. Linéarisation de la proposition relative en *que*

La proposition relative est bien connue pour illustrer la question de l'ordre sujet - verbe ou verbe - sujet. Dans cette sous-section, nous proposons d'abord deux exemples, puis une étude théorique, corroborée ensuite par une étude sur corpus.

Voici tout d'abord deux exemples de propositions relatives extraites du journal Le Monde. Dans la figure 16, le sujet de la relative, placé avant le verbe, a un complément de longueur 1, le verbe un complément de longueur 8. Le pronom relatif objet est à une distance 2 du verbe *fait imprimer* dont il sature la valence objet.

Figure 16 : exemple de linéarisation sujet - verbe

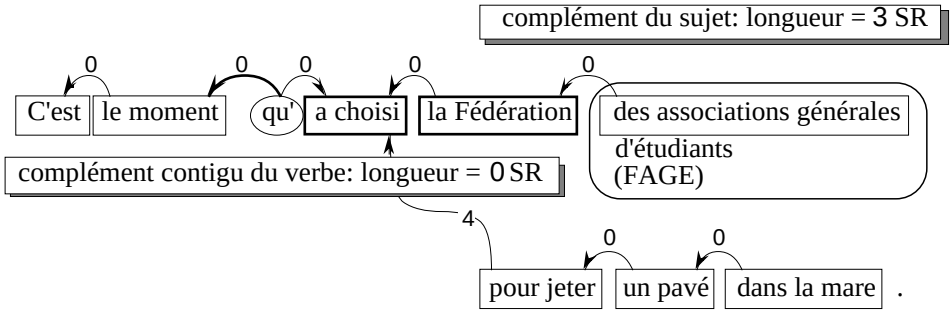


Dans la figure 17, le sujet de la relative, placé après le verbe, a un complément de longueur 3, le verbe n'a aucun complément contigu. Il a aussi des compléments non contigus, rejetés après les compléments du sujet, et qui ne font pas partie de la structure actancielle du verbe *a choisi*. Le

[9] Comme nous le rappelle Karl Popper, une hypothèse ne peut pas être prouvée, mais seulement corroborée ou falsifiée par des déductions dont les résultats (ici l'ordre des constituants) sont confrontés avec le matériau observable. Au lecteur de tenter de corroborer ou falsifier ces hypothèses sur d'autres types de corpus, d'autres langues, à l'écrit ou à l'oral.

pronom relatif objet est à une distance 0 du verbe *a choisi* dont il sature la valence objet.

Figure 17 : exemple de linéarisation verbe - sujet



Nous allons maintenant faire l'étude théorique de la linéarisation de la proposition relative en *que*, dans le cas où le sujet n'est pas un pronom sujet atone, mais un syntagme (qui peut donc avoir des compléments), et où le pronom relatif sature directement la valence objet du verbe de la relative, comme dans les figures 16 et 17, et non pas la valence objet d'un verbe infinitif ou conjugué dépendant du verbe de la relative, comme dans l'exemple suivant (voir son arbre linéarisé en annexe A.2.) :

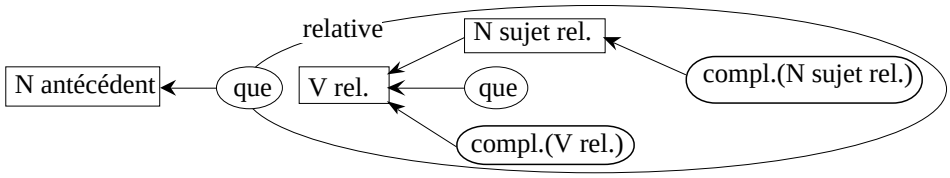
> *Ce document confirme l'importance des pertes* ***que*** *l'Etat sera amené à* ***supporter*** *sur les actifs détenus par le CDR.*

exemple où ***que*** sature la valence objet de l'infinitif ***supporter*** qui dépend de *sera amené* verbe conjugué de la relative.

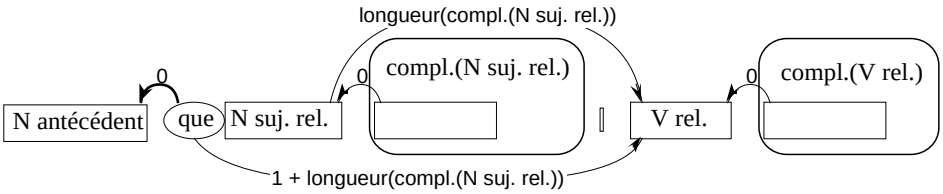
Dans la figure 18, seuls les actants du verbe sont présents, car ses circonstants peuvent être linéarisés après les compléments du sujet quand il est linéarisé après le verbe, comme dans l'exemple de la figure 17, et la linéarisation ne dépend pas de la présence éventuelle de ces circonstants.

Figure 18 :

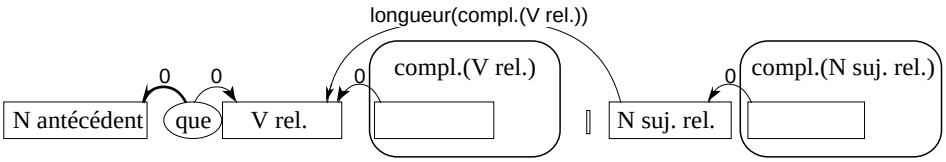
a) arbre de dépendance à transmettre :



b) linéarisation sujet puis verbe (cas de la figure 16)



c) linéarisation verbe puis sujet (cas de la figure 17)



Remarquons que la linéarisation verbe puis sujet place le verbe contigu à son objet *que*, le pronom relatif, et que la linéarisation sujet puis verbe éloigne le verbe de son objet *que*, d'autant plus que le sujet a beaucoup de compléments.

Pour chacune des deux linéarisations, calculons la somme (Σ) des longueurs des dépendances:

- linéarisation sujet puis verbe :

  Σ longueurs des dépendances = 1 + 2 * longueur (compléments (N sujet))

- linéarisation verbe puis sujet :

  Σ longueurs des dépendances = longueur (compléments (V))

La droite d'équilibre théorique entre les deux linéarisations possibles est la situation d'égalité de ces deux valeurs (voir ci-dessous figure 19) :

$$1 + 2 * \text{longueur (compléments (N sujet))} = \text{longueur (compléments (V))}$$

Pour que la linéarisation verbe puis sujet soit optimisée, il faut que la somme des longueurs des dépendances soit minimale :

$$\text{longueur (compléments (V))} \leq 1 + 2 * \text{longueur (compléments (N sujet))}$$

Les deux linéarisations placent en premier celui du sujet ou du verbe qui a peu ou aucun complément, et ensuite celui qui a beaucoup de compléments.

Les exemples des figures 16 et 17 corroborent cette propriété :

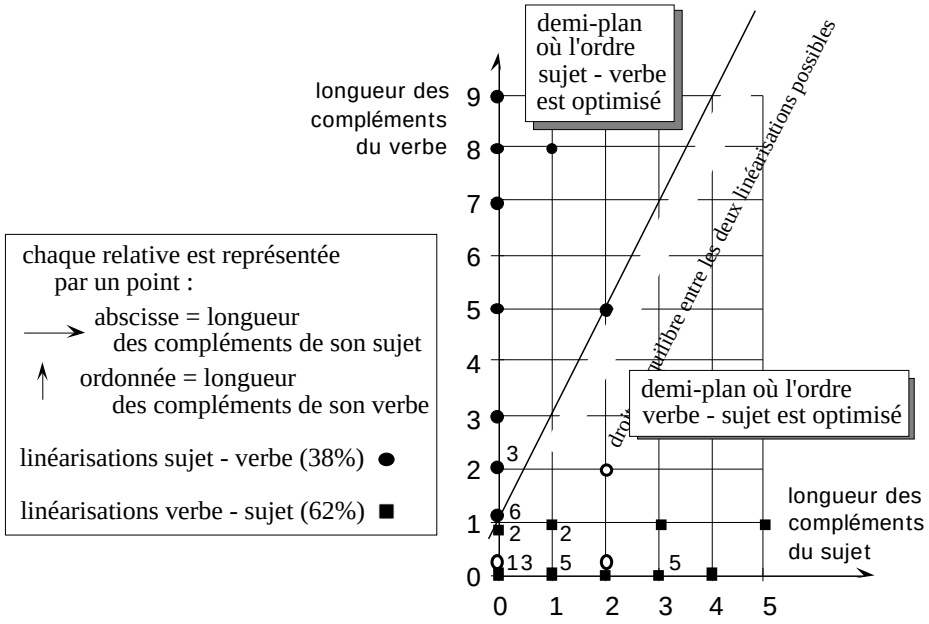
- figure 16 (ordre linéaire sujet puis verbe) : 8 SR > 1 + 2 * 1 SR
- figure 17 (ordre linéaire verbe puis sujet) : 0 SR < 1 + 2 * 3 SR

Nous présentons maintenant une validation sur corpus de ses propriétés.

Le corpus est composé d'articles du journal Le Monde qui ont servi à l'évaluation des tests du projet GRACE[10], action d'évaluation comparée des étiqueteurs du français. Ce corpus fait 46410 mots et 1886 phrases (fin de phrase par . ? ! ; :), et il a été étiqueté par Josette Lecomte (INALF), linguiste du comité de coordination du projet GRACE, ce qui nous a permis de localiser les *que* ou *qu'* pronoms relatifs. On compte alors :

65 relatives avec un pronom atone sujet (plus de la moitié)
61 relatives avec un syntagme sujet, parmi lesquelles:
9 relatives dont le pronom relatif sature la valence objet d'un infinitif
2 relatives - citations coupées, d'où une mesure impossible des longueurs
50 relatives restantes, parmi lesquelles:
19 linéarisations sujet puis verbe (38% des 50 relatives restantes)
31 linéarisations verbe puis sujet (62% des 50 relatives restantes)

Figure 19 : Repère cartésien longueur (compl. (V)) - longueur (compl. (N sujet))



Pour chacune de ces 50 relatives, dans la figure 19, nous avons placé un point qui la représente dans un repère cartésien orthonormé avec la longueur des compléments de son sujet en abscisse et la longueur des compléments de son verbe en ordonnée. Les points sont accompagnés de l'effectif (en italique) si celui-ci est supérieur à 1. De nombreux exemples d'arbre linéarisés de ces phrases sont donnés en annexe A.3.

[10] Voir la présentation et les résultats de ce projet d'évaluation comparative des étiqueteurs du français dans (Paroubek 98) et sur le site : http://www.limsi.fr/TLP/grace/ (février 99).

Remarquons que la linéarisation verbe puis sujet est la plus fréquente parmi les relatives où elle est possible (sujet syntagme).

Le tableau suivant fait le bilan statistique :

type de linéarisation	effectif	dans le demi-plan attendu	sur la droite d'équilibre	autres
sujet - verbe	19	9 (●)	7 (●)	3 (○)
verbe - sujet	31	29 (■)	2 (■)	0
totaux	50	38	9	3

On observe que l'étude théorique est validée et corroborée sur ce corpus pour 47 relatives sur 50. Trois relatives ne sont pas optimisées, ce sont des linéarisations sujet - verbe, marquées par un petit cercle blanc (○) sur la figure 19. Voici les trois phrases en question, suivies des longueurs des compléments du sujet et du verbe (voir leur arbre linéarisé en annexe A.3.5.) :

L'OTAN sera à l'Europe ce que l'Organisation des Etats américains (OEA) fut à l'Amérique latine dans les années 60 : un outil de coopération régionale certes, mais fonctionnant de manière inéquitable. (sujet : 2 SR, verbe : 2 SR)

De fait, le secteur financier français est l'un de ceux que le commissaire européen à la concurrence, Karel van Miert, connaît le mieux. (sujet : 2 SR, verbe : 0 SR)

Une gageure, lorsqu'on pense à la difficulté que la France a eu de faire adopter par ses quatorze partenaires de l'UE au sommet d'Amsterdam, le 18 juin, puis par les chefs des pays les plus industrialisés, à Denver le 22 juin, de simples déclarations relatives au Proche-Orient. (sujet : 0 SR, verbe : 0 SR)

Dans la première, une contrainte supplémentaire s'est exercée: la structure *X sera à Y ce que A fut à B* impose la linéarisation sujet - verbe. Les deux autres ne sont pas optimisées, probablement à cause une contrainte agissant sur un segment de taille supérieure à la phrase (voir les travaux de Nadine Lucas).

En ce qui concerne "La place du sujet nominal dans les relatives", on pourra consulter utilement (Fuchs 97), dont l'étude sur un numéro entier du journal Le Monde (157 relatives de tout type avec sujet postposable) fait de nombreuses hypothèses (thème-rhème, sémantique, longueur des groupes, ...). Nous citons ses observations sur la "longueur du constituant sujet" et sur la "longueur du constituant verbal" (mesures suggérées en syllabes et/ou en mots sans autre précision sur la métrique), page 153 :

*Si l'on considère la longueur **relative** du groupe sujet par rapport à celle du groupe verbal, on retrouve également la tendance énoncée plus haut : le groupe sujet a tendance à être postposé s'il est plus long que le groupe verbal, et a être antéposé s'il est plus court : c'est donc le plus long des deux groupes qui tend à être postposé.*

Ces observations corroborent globalement les nôtres, et nous pensons leur apporter le fondement théorique de la linéarisation optimisée sous la contrainte du moindre effort de mémoire.

3.2. Le processus de réception : phrase linéaire → arbre = mise en relation par l'intermédiaire de mémoires

Nous allons maintenant faire une hypothèse sur le processus par lequel, à la réception, l'arbre de dépendance est reconstruit à partir de l'ordre linéaire, en partant d'une définition dynamique de la mise en relation comme processus fondé sur l'utilisation de la mémoire : souvenir, puis oubli (une vue chronologique remplace la vue spatiale).

3.2.1. Définition de la "connexion" chez Tesnière

Souvenons-nous de la définition de la connexion chez Lucien Tesnière dans (Tesnière 59), page 11, § 3 (en fait la première page de Tesnière) :

> *Entre un mot et ses voisins, l'esprit aperçoit des **connexions**,*
> *dont l'ensemble forme la charpente de la phrase.*
> *Ces connexions ne sont indiquées par rien.*

Dans cette définition, Tesnière présente la connexion comme :

(1) un **processus** "*aperçoit*"
(2) un **processus** mental "*l'esprit*"
(3) un **processus** lié à la perception "*aperçoit*"
(4) un **processus** de calcul "*indiquées par rien*"

Notons que Tesnière définit sa connexion comme un processus exécuté par le lecteur - auditeur en situation de **réception**; ceci pose question sur l'ordre structural, qui peut être transformé en ordre linéaire en situation de **production**.

3.2.2. Hypothèse sur le processus de mise en relation

Partons de la connexion de Tesnière, définie comme processus de mise en relation en réception, et précisons-la comme un travail du récepteur (lecteur - auditeur) sur sa mémoire, un calcul présent sur la représentation présente d'événements passés (un état présent de la mémoire) :

en lisant - entendant un verbe :
a) le récepteur se souvient du sujet,
b) il relie le verbe au sujet,
c) puis il oublie ce sujet qui a trouvé son verbe, et n'en attend plus.

Le type de cette relation calculée par le récepteur est : le récepteur se souvient du sujet. Elle est orientée du verbe vers le sujet, du segment en cours de réception vers le segment dont il se souvient :

sujet <—— le récepteur se souvient —— verbe

Nous appelons cette relation intermédiaire : la "**souvenance**", pour la différencier de la dépendance. La "souvenance" est ensuite transformée en dépendance :

sujet —— dépend de ——> verbe

Pour se souvenir du sujet, le récepteur l'a auparavant mémorisé:

en lisant - entendant un syntagme nominal :
le récepteur le mémorise comme sujet potentiel

Nous proposons alors la définition suivante :

la **souvenance** est un processus et le résultat de ce processus, processus de mise en relation de 2 SR au cours de la réception par l'auditeur - lecteur, processus de calcul fondé sur sa mémoire, en deux temps distincts : temps de la réception du sujet potentiel, temps de la réception du verbe.

Remarquons que l'on trouve un concept analogue de processus en 2 temps dans (Grunig 93), page 15 :

> *si un adjectif* a *est suspendu depuis le temps t en attente d'un* n *et qu'il en rencontre un à un temps* t', *l'établissement en ce temps* t' *de la connexion entre* a *et* n *peut constituer le signal mettant un terme à la suspension.*

3.2.3. Le processus de mise en relation en deux temps modélisé et validé sur ordinateur

Nous allons maintenant décrire la modélisation sur ordinateur de ce processus de mise en relation, puis sa généralisation à tous types de relation, et enfin sa validation par son intégration dans un analyseur automatique et par l'analyse d'un corpus de taille importante. Remarquons que le modèle informatique n'est présenté ici que comme dispositif expérimental de validation du processus de mise en relation et non pas pour ses capacités opératoires - sous cet aspect, voir (Giguet-Vergne 97), (Vergne-Giguet 98) et (Giguet 98).

3.2.3.1. Présentation du modèle : cas de la mise en relation d'un verbe avec son sujet

Nous partons de l'hypothèse que cette relation de souvenance est établie par un processus en deux temps, par l'intermédiaire de mémoires spécialisées par type de relation.

Nous allons reprendre l'exemple de la relation entre un SR nominal sujet et un SR verbal, relation qui, au cours de la réception, sera établie en deux temps distincts et successifs :

- **temps 1** : un SR nominal est reçu, puis mémorisé comme sujet potentiel, en attente d'un SR verbal éventuel (voir figure 20, ci-dessous),
- **temps 2** : puis, si un SR verbal arrive (est reçu), il est relié au SR nominal en attente, qui alors n'attend plus de SR verbal, et est donc "oublié" dans la mémoire des sujets en attente de verbe (voir figure 21, ci-dessous).

Figure 20 : Temps **1** du processus de mise en relation entre deux SR

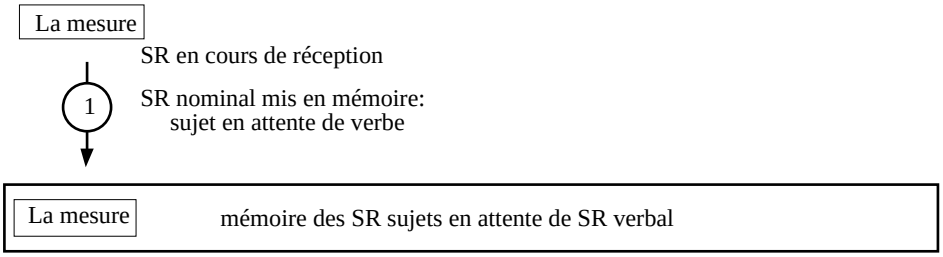
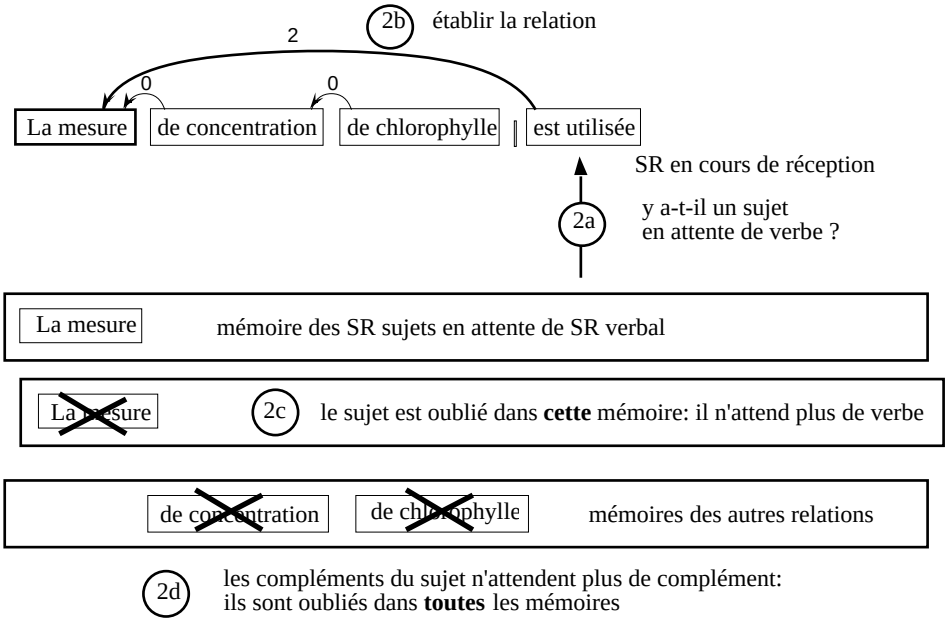


Figure 21 : Temps **2** du processus de mise en relation entre deux SR



Entre le temps 1 et le temps 2 de cette mise en relation sujet - verbe, les deux SR nominaux dépendant du sujet lui sont reliés par des processus identiques, mais par l'intermédiaire de la mémoire des régissants nominaux en attente de dépendants nominaux.

Le temps 2 comporte quatre opérations :

- 2a : la mémoire des sujets en attente de verbe est consultée,
- 2b : la relation est établie entre le verbe et le sujet en attente,
- 2c : le sujet n'attend plus de verbe : il est oublié de cette mémoire,
- 2d : tous les compléments du sujet sont oubliés de toutes les mémoires, car ils n'attendent plus de compléments : l'arrivée du verbe clôt définitivement la chaîne des compléments du sujet.

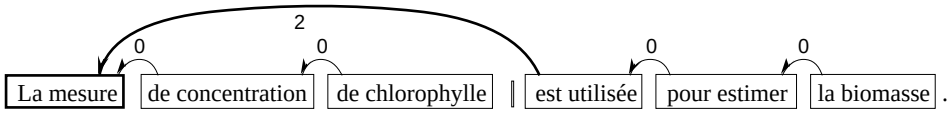
C'est la mise en relation en deux temps distincts, et plus particulièrement cette opération 2d, qui permet et concrétise l'interdépendance des différents processus de mise en relation en cours simultanément durant la réception. L'ensemble de ces processus interdépendants de mise en relation aboutit à l'arbre linéarisé des relations de souvenance de la figure 22 a).

3.2.3.2. Arbre de dépendance reconstruit par les processus de mise en relation

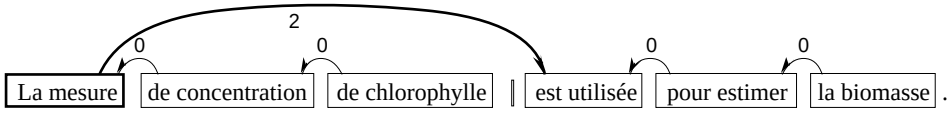
La figure 22 montre les trois étapes finales de la modélisation de la reconstruction de l'arbre de dépendance reçu par l'auditeur - lecteur. En particulier, les relations de dépendance sont restituées à partir des relations de souvenance, ce qui consiste à inverser les relations sujet - verbe. Enfin l'arbre est extrait de la linéarité, d'où la disparition de la métrique, et de toute trace écrite de la prosodie (ponctuation, groupes prosodiques, coupures prosodiques).

Figure 22 : Les trois étapes finales de la modélisation de la reconstruction de l'arbre de dépendance reçu par l'auditeur - lecteur

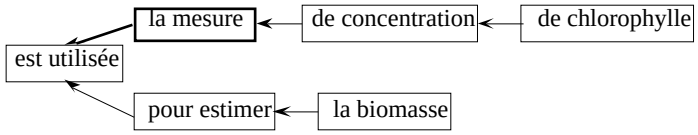
a) arbre linéarisé des relations de souvenance :



b) arbre de dépendance linéarisé, calculé à partir de l'arbre des relations de souvenance



c) arbre de dépendance reconstruit et reçu par l'auditeur - lecteur :



3.2.3.3. Généralisation du processus de mise en relation par l'intermédiaire de mémoires spécialisées pour chaque type de relation

Ce processus en deux temps est généralisé à l'aide de mémoires spécialisées, une pour chaque type de relation, ce qui organise et cloisonne les différents types d'attentes :

- mémoire des SR sujets	en attente d'un SR verbe
- mémoire des SR verbes transitifs	en attente d'un SR objet
- mémoire des SR	en attente d'un SR subordonné
- mémoire des SR	en attente d'un SR coordonné
- mémoire des SR antéposés	en attente d'un SR régissant
- mémoire des "*que*" pronom relatif	en attente d'un SR verbal transitif
- mémoire des "*ne*"	en attente de "*que*", "*ni*"
- ...	(les attendus arrivent ou n'arrivent pas)

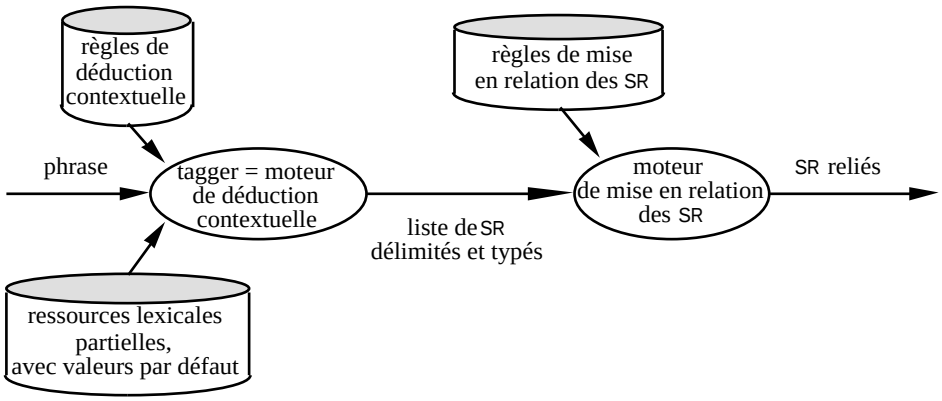
Chaque mise en relation **interagit** avec les autres mises en relation en cours en provoquant l'**oubli** des SR en attente entre les 2 SR reliés. C'est la seule interaction entre les différentes mémoires. Dans son état actuel, le modèle efface toutes les mémoires en fin de phrase, ce qui entraîne qu'aucune relation n'est établie entre deux phrases.

Notons qu'aucune hypothèse n'est faite explicitement sur les structures syntaxiques situées entre les deux segments reliés, ni sur la distance qui sépare ces deux segments, car seuls les processus sont explicités.

3.2.3.4. L'analyseur syntaxique automatique, cadre de la modélisation du processus de mise en relation

Ce processus est actuellement modélisé dans le cadre d'un analyseur syntaxique automatique qui produit l'arbre de dépendance sous la forme de la figure 22, avec les relations de coordination et d'antécédence.

Figure 23 : Architecture de l'analyseur syntaxique qui modélise le processus de mise en relation par l'intermédiaire de mémoires



Cet analyseur est décrit dans (Giguet-Vergne 97). Il est constitué de deux modules successifs : le premier, par une propagation de déductions contextuelles, produit la liste des SR délimités et typés, et le deuxième relie les SR en modélisant exactement et en généralisant le processus de mise en relation décrit ci-dessus, en l'implantant dans des règles déclaratives interprétées, avec un type de règle pour le temps 1 (mettre un SR en mémoire), et un autre type de règle pour le temps 2 (relier deux SR et oublier les SR situés entre eux)[11].

Pour les valider, ces concepts ont été confrontés avec des corpus variés en français (journal Le Monde, littérature, textes scientifiques) et de taille importante (corpus GRACE, de 650 000 mots), au moyen de l'analyseur syntaxique. L'efficacité opératoire est déjà importante[12] et valide le principe de la mise en relation en deux temps par l'intermédiaire de mémoires spécialisées par type de relation[13].

3.2.3.5. Rôle de l'ordinateur comme instrument de recherche en syntaxe

L'ordinateur étant une machine dont la fonction est d'exécuter un processus[14] sur des informations entrées dans sa mémoire (c'est-à-dire un enchaînement chronologique d'opérations sur ces informations, aboutissant à leur transformation), son utilisation comme outil de modélisation nous focalise en priorité sur l'explicitation des processus plutôt que sur l'explicitation des structures, et c'est ce qui nous a amené à concevoir ce processus de mise en relation, et à l'étendre ensuite à la définition de la souvenance.

L'ordinateur, comme machine chronologique, oblige à expliciter les processus à exécuter : des opérations ordonnées dans le temps, portant sur les informations traitées. Cette machine opère sur sa mémoire : elle mémorise puis efface des informations. Elle peut automatiser des chaînes de déductions et modéliser des processus hypothétiques, ce qui permet de vérifier leur faisabilité. Tel un microscope, ou un télescope, c'est un instrument au travers duquel on peut observer le matériau des corpus, observation filtrée par les concepts inscrits dans le programme : c'est donc

---

[11] Voir les corpus analysés ainsi sur internet à l'adresse : http://www.info.unicaen.fr/~giguet (février 99).

[12] Voir les résultats de ce projet d'évaluation comparative des étiqueteurs du français sur le site : http://www.limsi.fr/TLP/grace/ (février 99). Notre analyseur s'est placé en tête devant douze laboratoires français, suisses, allemands, québecois et huit entreprises françaises et américaines (dont IBM, Xerox, ATT Bell).

[13] Que nous apprend cette validation par analyse syntaxique automatique d'un corpus ? On constate la bonne capacité opératoire de l'analyseur, ainsi que la faisabilité de la généralisation des processus. On peut ensuite s'interroger sur la validité psycholinguistique de ces processus. Ceci est une conjecture ouverte : seules des expériences réalisées par des psycholinguistes pourront apporter un début de réponse.

[14] Ce processus est appelé algorithme, et est préenregistré dans la mémoire de l'ordinateur; sa forme concrète enregistrée constitue un programme informatique.

une machine d'aide à l'induction. Cette machine sert enfin à confronter des concepts avec ce matériau, et ainsi à corroborer ou falsifier des hypothèses sur ce matériau.

4. Conclusion

Dans cet article, nous avons tenté de présenter en un ensemble cohérent les deux processus de production et de réception qui permettent à un humain de transmettre un arbre de dépendance à un autre humain, associés aux deux formes structurelles que peut prendre une phrase : son ordre linéaire et son arbre de dépendance, arbre codé, compressé temporairement dans l'ordre linéaire; et nous avons mis en évidence les contraintes géométriques, informationnelles, chronologiques et mémorielles qui façonnent ces deux processus et ces deux structures.

Références bibliographiques

Abney, Steven (1996), "Part-Of-Speech Tagging and Partial Parsing", in Ken Church and Steve Young and Gerrit Bloothooft, editors, *An Elsnet Book*, Corpus-Based Methods in Language and Speech, Kluwer Academic, Dordrecht.

Chomsky, Noam (1969), *Structures syntaxiques*, Point Seuil, Paris.

Chomsky, Noam (1971), *Aspects de la théorie syntaxique*, Seuil, Paris.

Chomsky, Noam (1980), *Essais sur la forme et le sens*, Seuil, Paris.

Déjean, Hervé (1998a), "Inférences automatiques de contextes distributionnels", in *Actes de la conférence TALN'98*, Paris.

Déjean, Hervé (1998b), *Apprentissage des structures syntaxiques des langues*. Thèse de doctorat en Informatique de l'université de Caen.

Giguet, Emmanuel, et Vergne, Jacques (1997), "From part of speech tagging to memory-based deep syntactic analysis", in *International Workshop on Parsing Technologies 1997 Proceedings*, Boston.

Giguet, Emmanuel (1998), *Méthode pour l'analyse automatique de structures formelle sur document multilingues*, Thèse de doctorat en Informatique de l'université de Caen.

Fuchs, Catherine (1997), "La place du sujet nominal dans les relatives", in *La place du sujet en français contemporain*, pp. 135-178, Duculot, Louvain.

Gaifman, Haïm (1965), "Dependency Systems and Phrase Structure Systems", in *Information and Control* n°8, pp. 304-337.

Grunig, Blanche-Noëlle (1993), "Charges mémorielles et prédictions syntaxiques", in *Cahiers de Grammaire* n°18, ERSS, Toulouse, pp. 13-29.

Hays, D.G. (1964), "Dependency theory", in *Language* n°40, pp. 511-525.

Kahane, Sylvain (1997), "Bubble trees and syntactic representations", in *Proc. 5th Meeting of the Mathematics of Language (MOL5)*, Saarbrücken, pp. 70-76.

Le Goffic, Pierre (1993), *Grammaire de la Phrase Française*, Hachette, Paris.

Lucas, Nadine (1993a), "Syntaxe du paragraphe dans les textes scientifiques en japonais et en français", in *Actes du colloque international : Parcours linguistiques de discours spécialisés*, sous la direction de Sophie Moirand, éditions Peter Lang, Berne-Paris, pp. 249-261.

Lucas N., K. Nishina, T. Akiba et K.G. Suresh (1993b), *Discourse analysis of scientific textbooks in Japanese : a tool for producing automatic summaries*, Department of Computer Science, Tôkyô Institute of Technology, Tôkyô.

Lucas, Nadine (1995), "Le style scientifique en japonais et en français", in *Japon Pluriel, actes du 1er colloque de la Société française des études japonaises*, éditions Picquier, Arles, pp. 393-402.

Mel'cuk, Igor (1988), *Dependency syntax : theory and practice*, State University Press NY, Albany, NY.

Paroubek, Patrick (1998), "Experience in Grace tagging evaluation", in *First International Conference on Language Resources and Evaluation proceedings*, Granada.

Portine, Henri (1992), "Ordre structural et ordre linéaire chez Tesnière", in Actes du Colloque International *Lucien Tesnière aujourd'hui*, CNRS URA 1164, Université de Rouen, pp. 119-127.

Stevens, Peter (1978), *Les formes dans la nature*, Seuil, Paris.

Tesnière, Lucien (1953), *Esquisse d'une syntaxe structurale*, Klincksieck, Paris.

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

Vergne, Jacques (1995a), "Les cadres théoriques des Traitements Automatiques des Langues syntaxiques : quelle adéquation linguistique et algorithmique? une étude et une alternative", in *Actes de "TALN'95" Conférence sur le Traitement Automatique du Langage Naturel*, Marseille, pp. 24-33.

Vergne, Jacques (1995b), "Esquisse d'une syntaxe des langues concrètes - Application à l'analyse syntaxique automatique", in *Les cahiers du GREYC*, numéro 11, université de Caen, 180 pages.

Vergne, Jacques, et Giguet, Emmanuel (1998), "Regards Théoriques sur le “Tagging”", in *Actes de la cinquième conférence annuelle : Le Traitement Automatique des Langues Naturelles*, TALN'98, Paris, pp. 22-31.

Zipf, George Kingsley (1949, réédition 1966), *Human Behavior and the Principle of Least Effort*, Harper, New York.

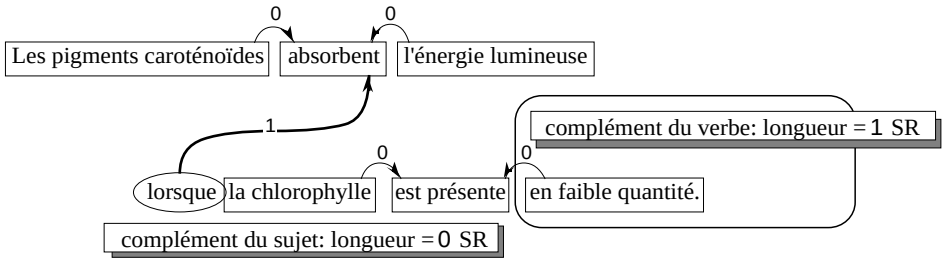
Annexes : Exemples d'arbres linéarisés

Dans ces annexes, nous proposons au lecteur des cas concrets et variés d'arbres linéarisés de phrases extraites du corpus (journal Le Monde), pour mettre les concepts à l'épreuve du matériau linguistique.

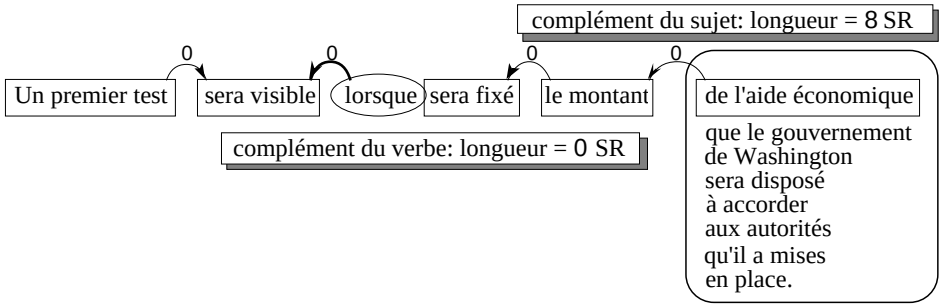
A.1. Deux linéarisations de propositions subordonnées circonstancielles avec *lorsque* (optimisées)

Du sujet et du verbe, celui qui a le moins de compléments est linéarisé le premier.

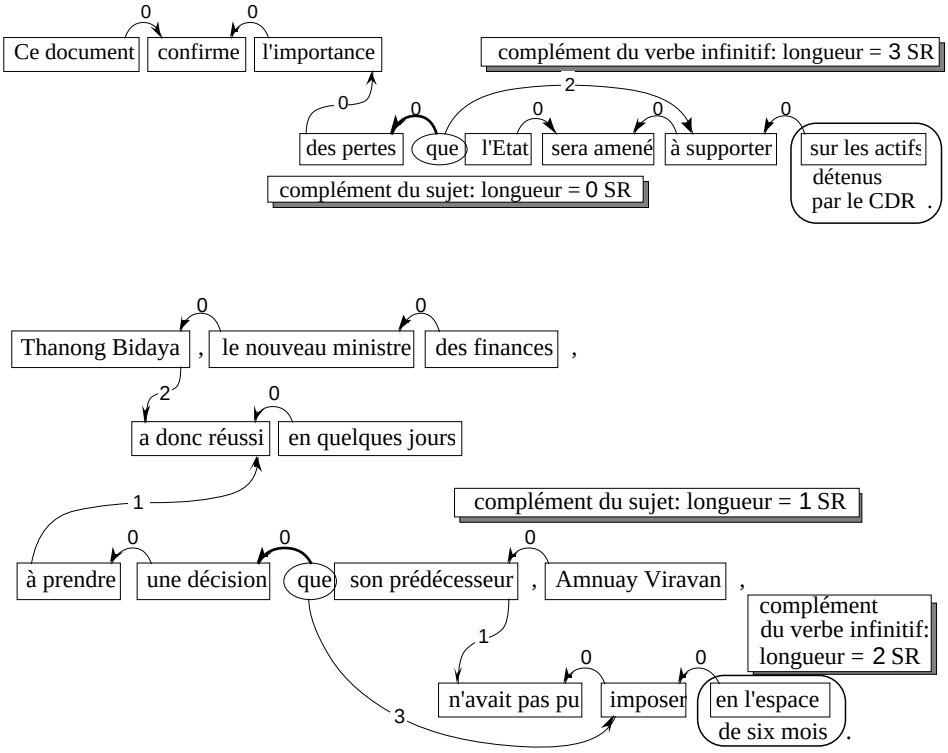
Exemple de linéarisation sujet - verbe (0 < 1) :



Exemple de linéarisation verbe - sujet (0 < 8) :



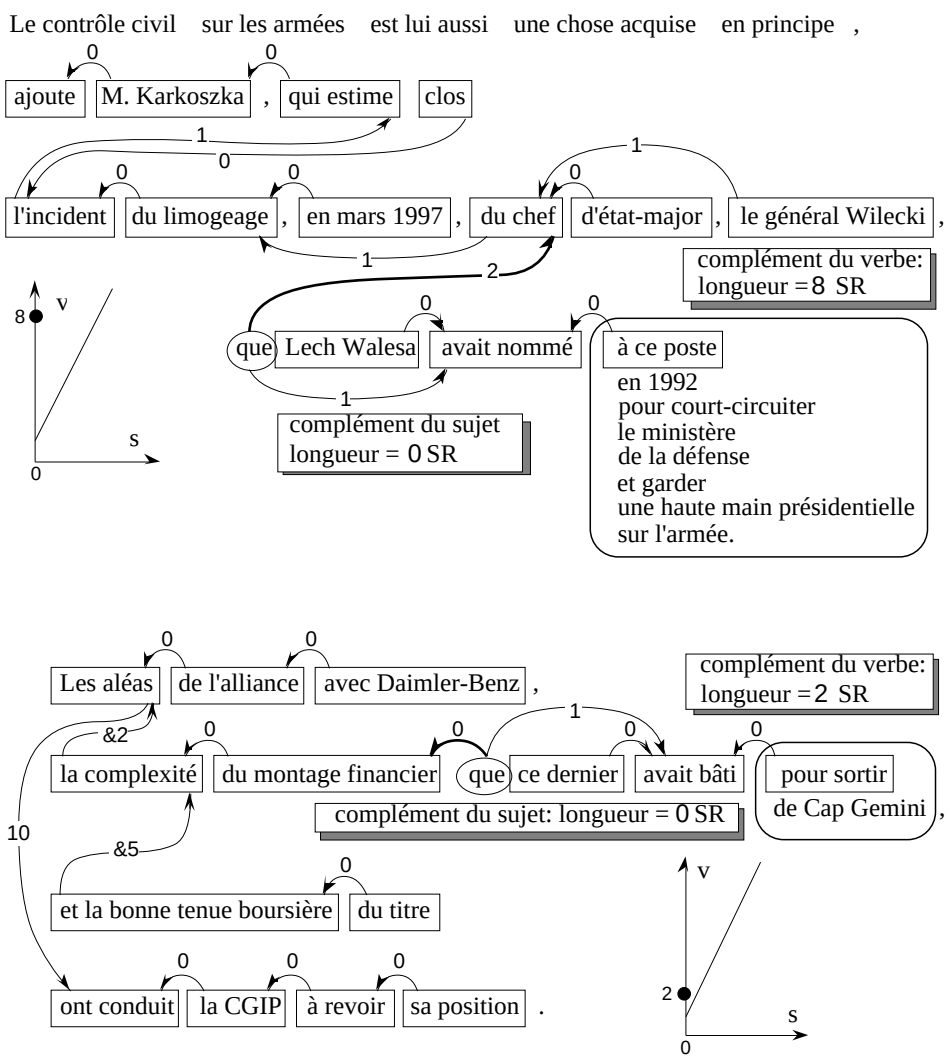
A.2. Linéarisation sujet - verbe de propositions subordonnées relatives où *que* est objet d'un infinitif dépendant du verbe de la relative



A.3. Linéarisation de la proposition relative en *que*

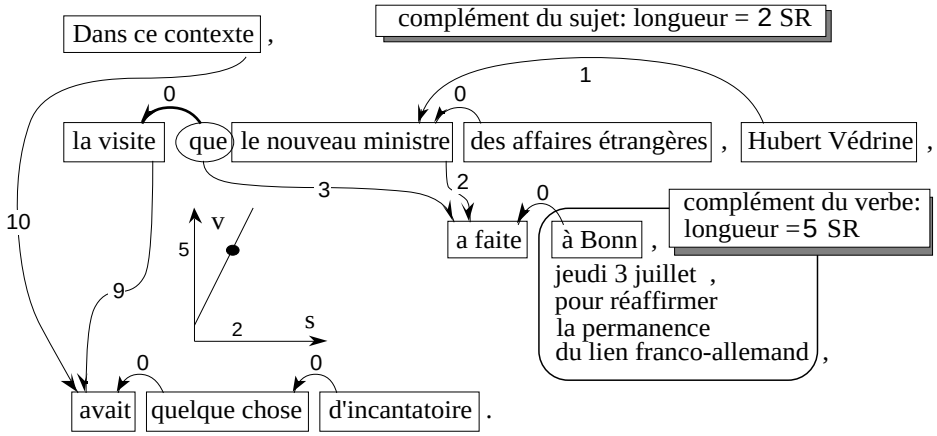
Dans les exemples suivants, nous présentons quelques relatives extraites du corpus des 50 relatives étudiées en 3.1.5. avec leurs coordonnées dans le repère cartésien de la figure 19, la longueur des compléments de son sujet en abscisse et la longueur des compléments de son verbe en ordonnée, en allant des linéarisations sujet - verbe vers les linéarisations verbe - sujet, en passant par des linéarisations équilibrées (pour chaque exemple, on rappelle sa position sur une réduction du repère de la figure 19).

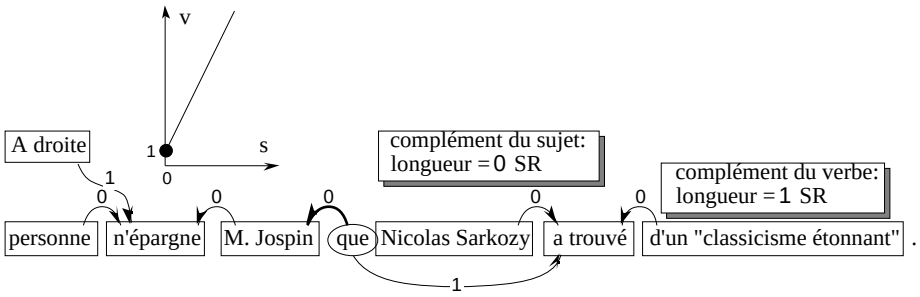
A.3.1. Linéarisations sujet - verbe dans le demi-plan sujet - verbe (optimisées)



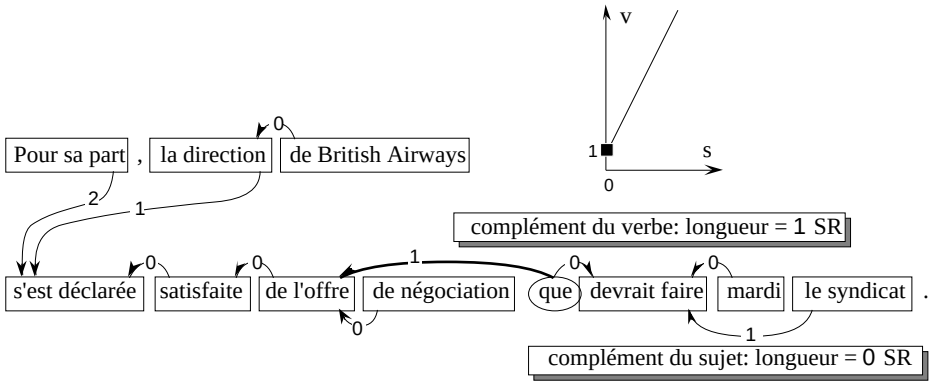
NB : les relations de coordination sont marquées par le signe & devant la longueur de la coordination.

A.3.2. Linéarisations sujet - verbe sur la droite d'équilibre (donc optimisées)

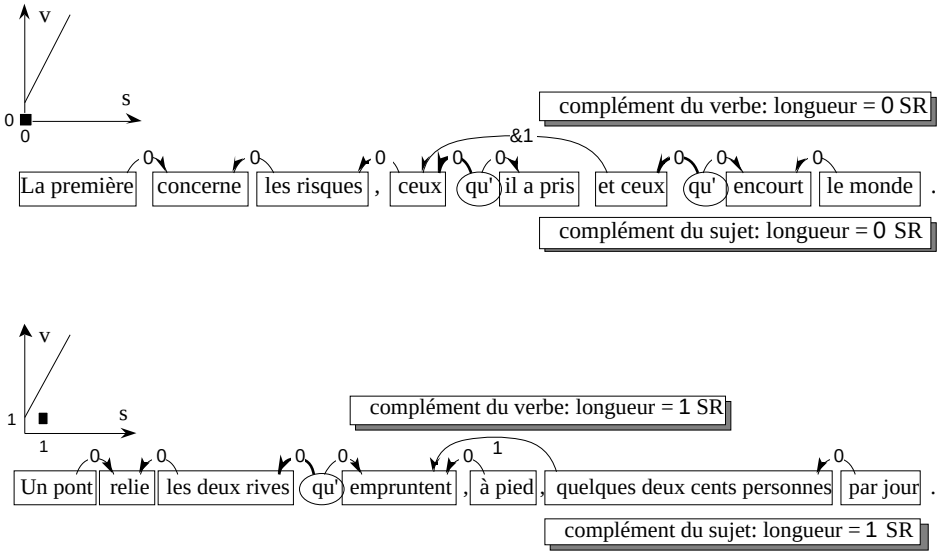


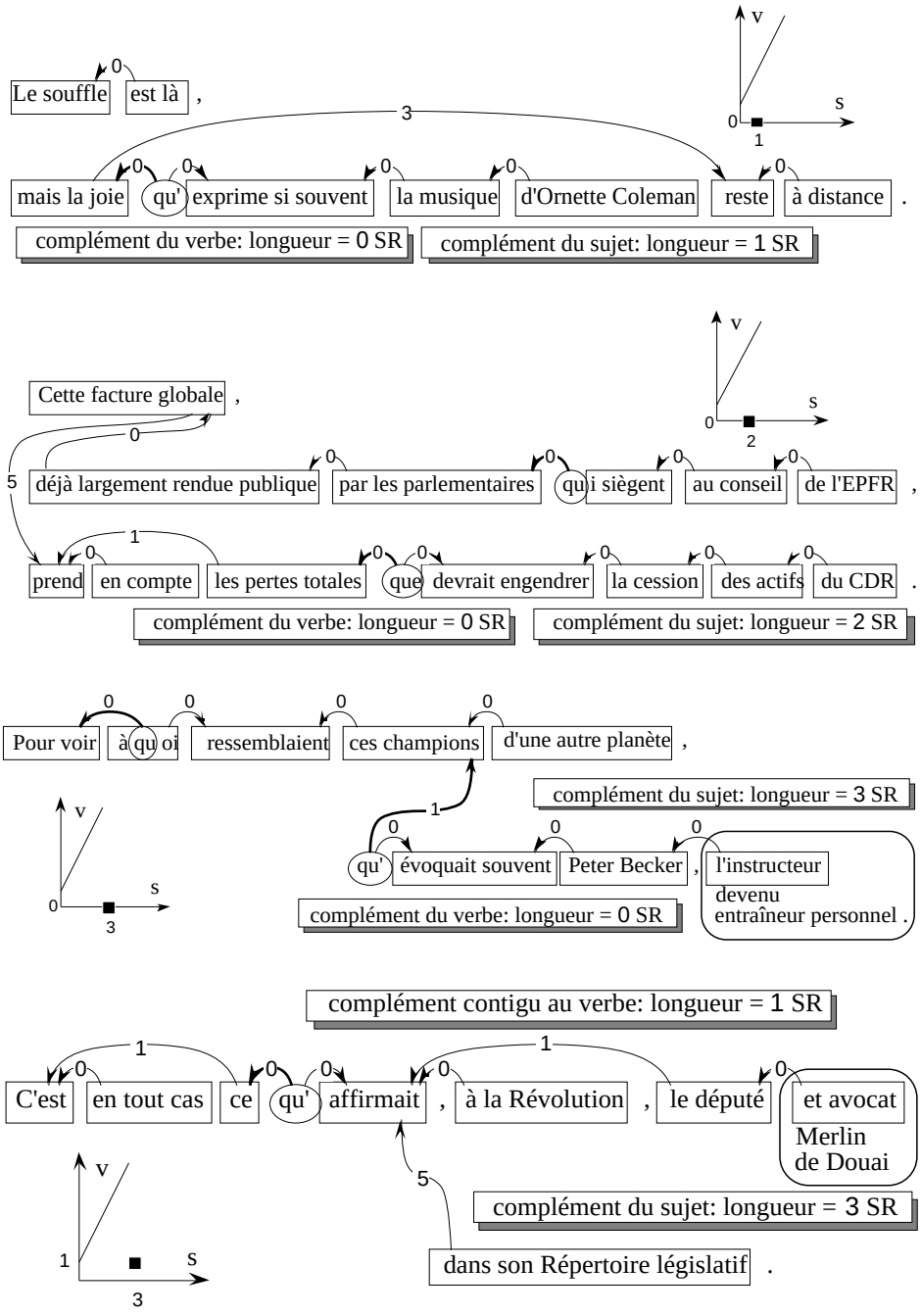


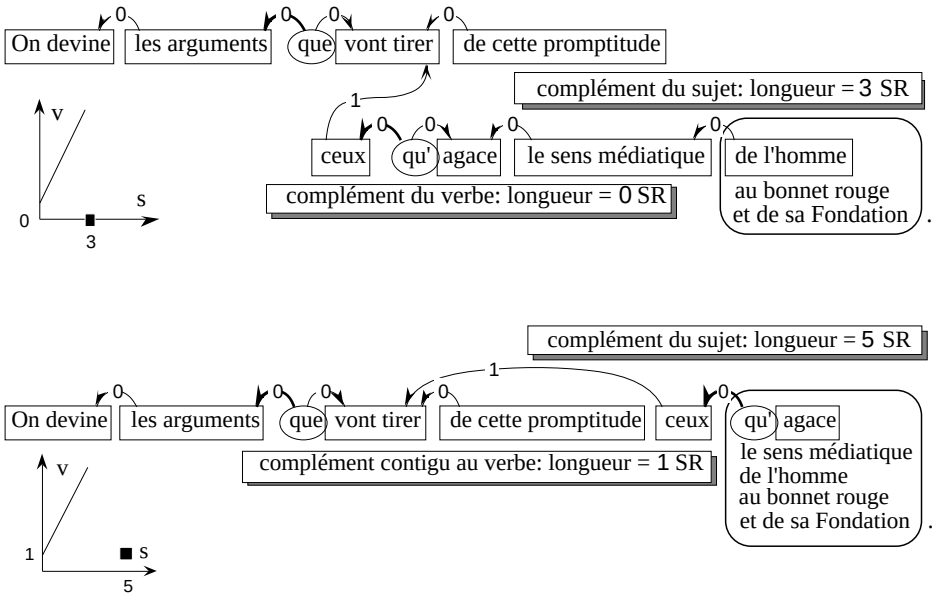
A.3.3. Linéarisation verbe - sujet sur la droite d'équilibre (donc optimisées)



A.3.4. Linéarisations verbe - sujet dans le demi-plan verbe - sujet (optimisées)



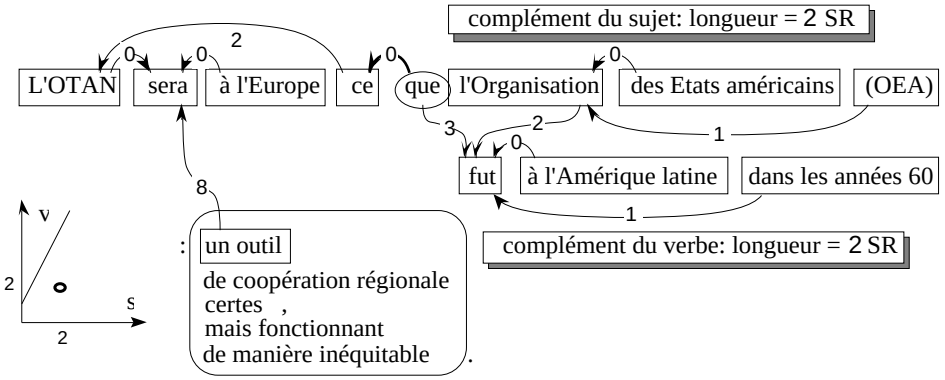

Le souffle
est là ,
mais la joie
qu'
exprime si souvent
la musique
d'Ornette Coleman
reste
à distance .
complément du verbe: longueur = 0 SR
complément du sujet: longueur = 1 SR
Cette facture globale ,
déjà largement rendue publique
par les parlementaires
qui siègent
au conseil
de l'EPFR ,
prend
en compte
les pertes totales
que
devrait engendrer
la cession
des actifs
du CDR .
complément du verbe: longueur = 0 SR
complément du sujet: longueur = 2 SR
Pour voir
à qui
ressemblaient
ces champions
d'une autre planète ,
complément du sujet: longueur = 3 SR
qu'
évoquait souvent
Peter Becker ,
l'instructeur devenu entraîneur personnel .
complément du verbe: longueur = 0 SR
complément contigu au verbe: longueur = 1 SR
C'est
en tout cas
ce
qu'
affirmait
, à la Révolution ,
le député
et avocat Merlin de Douai
complément du sujet: longueur = 3 SR
dans son Répertoire législatif .



A.3.5. Linéarisations sujet - verbe dans le demi-plan verbe - sujet (non optimisées)

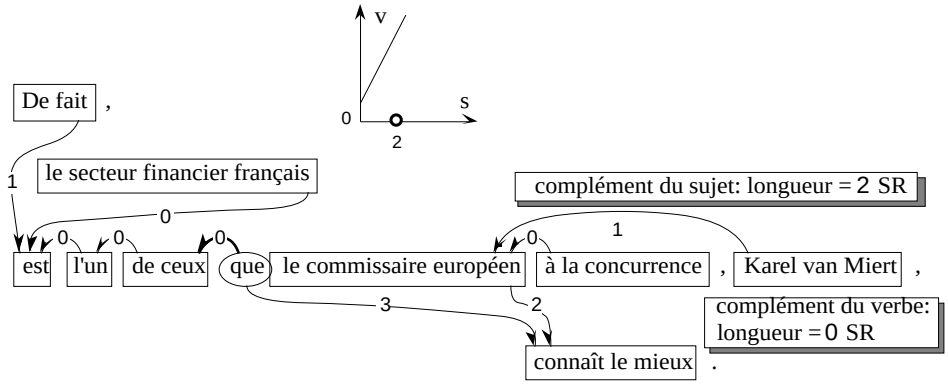
Voici les trois phrases du corpus dont la linéarisation n'est pas optimisée (voir aussi en fin de 3.1.5.) et pour lesquelles nous invitons le lecteur à faire des hypothèses sur les contraintes supplémentaires qui se sont exercées. Comme ces contraintes sont éventuellement au niveau supérieur, nous citons chaque phrase dans son paragraphe complet.

> *L'Europe de Vancouver à Vladivostok, appelée de ses vœux au sortir de la guerre froide par James Baker, le secrétaire d'Etat de George Bush, se met en place. Sa capitale est Washington. Le cadre institutionnel est fourni par l'OTAN qui est débarrassé des nécessités d'une défense collective, accentue son rôle politique. Elle devient ainsi l'instrument de l'influence américaine en Europe. L'OTAN sera à l'Europe ce* ***que l'Organisation des Etats américains (OEA) fut à l'Amérique latine dans les années 60*** *: un outil de coopération régionale certes, mais fonctionnant de manière inéquitable. Les priorités et orientations sont définies par l'acteur principal, les autres ayant pour tâche d'acquiescer et d'appliquer. Au nom de la défense des intérêts supérieurs de la collectivité, une politique en tous points conforme à celle du pays Leader se met en place.*



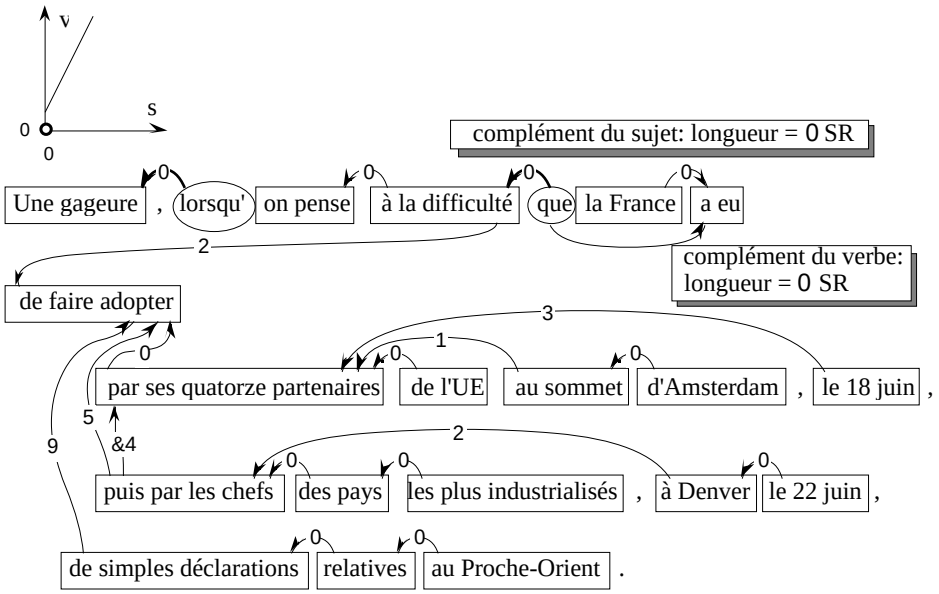
*

"Ce qui me guidera, (...) ce sera aussi le coût que représentent pour l'Etat et donc pour les contribuables les recapitalisations fréquentes", a-t-il précisé, rappelant en outre que l'Etat "ne peut pas toujours le faire dans le droit communautaire face à la Commission". De fait, le secteur financier français est l'un de ceux ***que le commissaire européen à la concurrence, Karel van Miert, connaît le mieux.*** *Il a depuis plusieurs années trois dossiers sur son bureau : outre celui du Gan, celui du Crédit lyonnais bien sûr, mais également celui de la Marseillaise de Crédit, trois établissements que les prédécesseurs de Lionel Jospin se sont engagés à privatiser.*



*

*M. Arafat a suggéré lundi à Paris l'utilisation du levier économique européen envers Israël. Une gageure, lorsqu'on pense à la difficulté **que la France a eu** de faire adopter par ses quatorze partenaires de l'UE au sommet d'Amsterdam, le 18 juin, puis par les chefs des pays les plus industrialisés, à Denver le 22 juin, de simples déclarations relatives au Proche-Orient. Le conseil européen d'Amsterdam invitait "les peuples du Proche-Orient à s'associer aux peuples d'Europe pour bâtir un avenir harmonieux" et engageait "les dirigeants israéliens et palestiniens" à faire avancer les choses.*



*