Grammaires de dépendance formelles et théorie Sens-Texte

Sylvain Kahane

Lattice, Université Paris 7, UFRL
Case 7003, 2, place Jussieu, 75251 Paris cedex 5
sk@ccr.jussieu.fr

Résumé – Abstract

On appelle grammaire de dépendance toute grammaire formelle qui manipule comme représentations syntaxiques des structures de dépendance. Le but de ce cours est de présenter à la fois les grammaires de dépendance (formalismes et algorithmes de synthèse et d'analyse) et la théorie Sens-Texte, une théorie linguistique riche et pourtant méconnue, dans laquelle la dépendance joue un rôle crucial et qui sert de base théorique à plusieurs grammaires de dépendance.

We call dependency grammar every grammar which handles dependency structures as syntactic representations. The aim of this course is to present both dependency grammars (formalisms, analysis and synthesis algorithms) and the Meaning-Text theory, a rich but nevertheless unrecognized linguistic theory, in which dependency plays a crucial role and which serves as theoretical basis of several dependency grammars.

1 Introduction

La représentation syntaxique d'une phrase par un arbre de dépendance est certainement plus ancienne que la représentation par un arbre syntagmatique. L'usage des dépendances remonte à l'antiquité. Les grammairiens arabes du 8ième siècle, comme Sibawaih, distinguaient déjà gouverneur et gouverné en syntaxe et utilisait cette distinction pour formuler des règles d'ordre des mots ou de rection (Owens 1988:79-81). On trouve des représentations de structures de dépendance dans des grammaires du 19éme siècle (Weber 1992:13). La première théorie linguistique basée sur la dépendance est incontestablement celle de Tesnière (1934, 1959), sans minimiser des travaux précurseurs, comme les représentations "moléculaires" de Jespersen (1924) ou la syntaxe du russe de Peškovskij (1934). Peu après, Hays (1960, 1964) développait la première grammaire de dépendance, tandis que Gaifman (1965) établissait les liens entre les grammaires de dépendance de Hays, les grammaires catégorielles de Bar-Hillel et les grammaires de réécriture de Chomsky. A l'exception de la grammaire de Robinson (1970), les grammaires de dépendance se sont ensuite surtout développées en Europe, notamment autour de Sgall et Hajičová à Prague (Sgall *et al.* 1986) et de Mel'čuk à Moscou (Mel'čuk 1974, 1988a), ainsi qu'en Allemagne (cf., par ex., la classique grammaire de l'allemand de Engel 1992) et au Royaume Uni autour de Anderson (1971) et Hudson (1990), la France restant curieusement à l'écart.

La représentation syntaxique d'une phrase par une structure syntagmatique, quant à elle, ne s'est développée qu'à partir de Bloomfield (1933) et des travaux des distributionnalistes. L'engouement formidable pour les grammaires génératives-transformationnelles de Chomsky (1957, 1965) dans les années 60-70 a retardé l'essor des grammaires de dépendance. Pourtant,

depuis la fin des années 70 et l'avènement de la Syntaxe X-barre, la plupart des modèles linguistiques issus de la mouvance chomskienne (GB/PP/MP, LFG, G/HPSG) ont introduit l'usage de la dépendance syntaxique sous des formes plus ou moins cachées (fonctions syntaxiques, constituants avec tête, cadre de sous-catégorisation, c-commande). De leur côté, les grammaires complètement lexicalisées comme les Grammaires Catégorielles ou TAG (Joshi 1987), en dérivant une phrase par combinaison de structures élémentaires associées aux mots de la phrase, construisent, par un effet de bord, des structures de dépendances.

Le retour de la dépendance au premier plan, au cours des années 80, est dû à deux facteurs principaux : le retour en grâce du lexique d'une part et de la sémantique d'autre part. Pour le lexique, grammaires de dépendance, en mettant la lexie au centre de la structure syntaxique, permettent d'exprimer simplement les relations lexicales comme la valence et le régime (ou sous-catégorisation). Pour la sémantique, les structures de dépendance, en permettant de dissocier l'ordre des mots et la structure syntaxique proprement dite, se rapprochent davantage d'une représentation sémantique que ne le fait une structure syntagmatique. Mieux encore, les relations sémantiques prédicat-argument, parfois appelées dépendances sémantiques, bien que devant être distinguées des dépendances syntaxiques, coïncident en partie avec celles-ci (cf. Mel'čuk 1988b, Kahane & Mel'čuk 1999).

Enfin, les grammaires de dépendance prouvent à l'heure actuelle leur bonne adéquation au traitement automatique des langues. Citons deux systèmes d'envergure développés en France : le générateur de texte développé à LexiQuest par Coch (1996) et intégré au système MultiMétéo (Coch 1998) de génération de bulletins météo multilingues et l'analyseur en flux de Vergne (2000) qui a remporté l'action Grace, portant sur l'évaluation des étiqueteurs pour le français. Pour d'autres travaux, on pourra également se reporter aux actes du dernier atelier sur le traitement automatique par des grammaires basées sur la dépendance (Kahane & Polguère 1998), au numéro spécial de la revue TAL sur les grammaires de dépendance (Kahane 2000c) et au portail officiel des grammaires de dépendance (http://ufal.mff.cuni.cz/dg.html).

Comme annoncé dans le titre, cet exposé est consacré aux grammaires de dépendance en général et à la théorie Sens-Texte en particulier. Dans la Section 2, nous tenterons de caractériser les notions de base de dépendance syntaxique et de fonction syntaxique et nous présenterons les premières grammaires de dépendance. La Section 3 sera consacrée à la théorie Sens-Texte [TST], probablement la plus achevée des théories linguistiques basées sur la dépendance. Dans la Section 4, nous ferons le lien entre la TST, dont les règles servent à mettre en correspondance des structures, et les grammaires génératives, dont les règles servent à générer des structures, et nous proposerons une grammaire de dépendance basée sur les principes théoriques de la TST, mais utilisant un formalisme d'unification. Dans la Section 5, nous nous pencherons sur les techniques de base pour l'analyse avec une grammaire de dépendance.

Je souhaite insister sur le fait que cet exposé n'est pas un survol impartial du domaine des grammaires de dépendance, loin de là. Il s'agit clairement d'un point de vue personnel sur la question, enrichi, je l'espère, de mes nombreuses discussions avec Igor Mel'čuk sur les fondements de la théorie Sens-Texte et sur le rôle de la dépendance en linguistique. J'en profite pour le remercier chaleureusement pour ses nombreuses remarques sur la première version de ce texte.

2 Arbres et grammaires de dépendance

Dans la Section 2.1, nous tenterons de caractériser la notion de dépendance syntaxique. A travers divers exemples, nous exposerons les points qui font l'unanimité entre les différentes théories et ceux qui posent problème (l'auxiliaire, la conjonction de subordination, le déterminant, le pronom relatif, la coordination, …). Nous nous intéresserons aux différentes façons d'encoder la dépendance et en particulier à l'équivalence entre arbres de dépendance et arbres syntagmatiques avec têtes. Dans la Section 2.2, nous nous intéresserons à la notion de

fonction syntaxique qui est indissociable de celle de dépendance syntaxique. Enfin, dans la Section 2.3, nous présenterons les premières grammaires de dépendance (Hays 1964, Gaifman 1965) et leur lien avec les grammaires catégorielles et les grammaires de réécriture hors-contexte.

2.1 Caractérisation de la notion de dépendance syntaxique

La quasi-totalité des théories linguistiques s'accordent sur le fait que, au-delà de la question du sens, les mots d'une phrase obéissent à un système d'organisation relativement rigide, qu'on appellera la *structure syntaxique* de la phrase. Il existe deux grands paradigmes pour représenter cette structure : soit décrire la façon dont les mots peuvent être groupés en des paquets de plus en plus gros (ce qui a donné les *structures syntagmatiques*), soit expliquer la façon dont les mots, par leur présence, dépendent les uns des autres (ce qui a donné les *structures de dépendance*). Comme on le verra, les deux paradigmes s'opposent davantage sur la façon de présenter l'organisation syntaxique que sur la nature même de cette organisation.[1]

Considérer qu'un *arbre de dépendance syntaxique* peut rendre compte des propriétés syntaxiques d'une phrase, c'est considérer que dans une phrase, la présence de chaque mot (sa nature et sa position) est légitimée par la présence d'un autre mot (son *gouverneur syntaxique*), à l'exception d'un mot, le mot principal associé au sommet de l'arbre syntaxique. La dépendance syntaxique est donc une dépendance entre mots (Figure 1 à gauche).

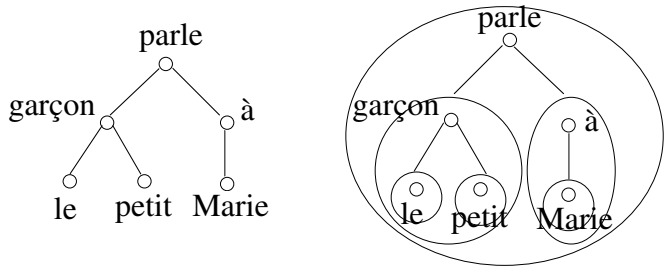

Figure 1 : Arbre de dépendance et arbre à la Gladkij pour *Le petit garçon parle à Marie*

Cette présentation de la structure syntaxique de la phrase est souvent mal acceptée de ceux qui voient plutôt des relations entre des mots et des groupes de mots que des relations entre des mots seulement. Précisons ce point. Quand un mot *x* légitime la présence d'un mot *y* (c'est-à-dire quand *x* gouverne *y*), en fait, par transitivité, *x* légitime également la présence des mots légitimés par *y* et des mots légitimés par ceux-ci. En conséquence, *x* légitime non seulement la présence de *y*, mais la présence d'un groupe de mot, qu'on appelle la *projection* de *y*. On peut donc présenter la structure de dépendance non pas comme des dépendances entre mots, mais comme des dépendances entre des mots et des groupes de mots (à l'intérieur desquels il y a à nouveau des dépendances entre mots et groupes de mots. Cette structure de dépendance entre des mots et des groupes peut être représentée par une structure que nous appellerons *un "arbre" à la Gladkij* (Figure 1 à droite) (Gladkij 1968, Kahane 1997). A l'intérieur de chaque groupe

---

[1] Je parle ici des représentations elles-mêmes. Il existe bien sûr des oppositions plus fondamentales qui ont conduit les uns où les autres à développer telle ou telle manière de présenter les choses. En particulier, comme je l'ai déjà dit, la grammaire syntagmatique est née d'une vision purement orienté vers l'analyse (à partir du texte), le distributionnalisme, et d'un rejet presque absolu des différences lexicales (Gross 1975) et des questions de sémantique.

ainsi considéré, il y a un mot qui n'appartient à aucun sous-groupe et qu'on appelle la *tête*[2]. On peut aussi représenter l'arbre à la Gladkij par une *structure syntagmatique avec tête lexicale*, c'est-à-dire une structure syntagmatique traditionnelle[3] où chaque constituant possède un sous-constituant tête qui est un mot (voir Figure 2 où la structure syntagmatique est représentée, à gauche, par un enchâssement de groupe et, à droite, par un arbre non étiqueté formellement équivalent ; dans les deux cas, le sous-constituant tête est indiqué par l'étiquette T). Le fait de considérer pour chaque constituant une tête n'est pas nouveau (cf. par ex. Pittman 1948). Ceci est devenu monnaie courante depuis la Syntaxe X-barre (Jackendoff 1977).

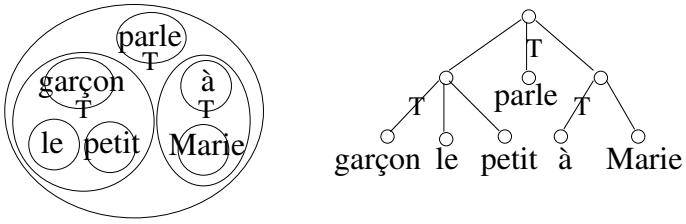

Figure 2 : Arbres syntagmatiques avec têtes
pour *Le petit garçon parle à Marie*

Les structures syntagmatiques avec tête et les arbres de dépendance (entre mots) sont formellement équivalents (Gladkij 1966, Robinson 1970). On a vu comment on passe d'un arbre de dépendance à une structure syntagmatique avec tête en introduisant les groupes obtenus par transitivation de la relation de dépendance. Inversement, on passe d'un arbre syntagmatique avec tête lexicale à un arbre de dépendance en ne représentant plus les groupes et en reliant le gouverneur d'un groupe directement avec la tête de ce groupe.[4]

Après nous être attaché aux différentes façons de représenter formellement la dépendance, nous allons aborder la question de la caractérisation théorique de la dépendance. Tesnière lui-même ne caractérise pas clairement la dépendance. Mel'čuk 1988a propose, à la suite de Garde 1977, une tentative de caractérisation directement en terme de dépendance entre mots. Du fait de l'équivalence entre arbres de dépendance et structure syntagmatique avec tête, il est également possible de caractériser la dépendance en caractérisant le constituant, puis la tête d'un constituant. Nous ne nous attarderons pas sur la façon d'identifier les constituants, mais sur la façon d'identifier la tête d'un constituant. Concernant les différentes définitions possibles de la tête et les cas litigieux, citons tout particulièrement le travail de Zwicky 1985. Nous adopterons ici la définition suivante (Mel'čuk 1988a) :

---

[2] Nous distinguons clairement les termes *tête* et *gouverneur*. Le *gouverneur x* d'un mot *y* (ou d'un goupe G) est le mot qui légitimise la présence de *y* (ou de G). Il n'appartient pas à G. Par contre, la *tête y* du groupe G est un mot de G qui légitimise, par transitivité, la présence des autres mots de G.

[3] L'arbre de dépendance n'encode pas l'ordre linéaire des mots. Bien que ce ne soit pas l'usage, on parle donc ici d'un arbre syntagmatique non ordonné.

[4] On peut également obtenir un arbre de dépendance à partir d'une structure syntagmatique avec tête où on autorise le sous-constituant tête à être un groupe de mots. En d'autres termes, on autorise en fait un mot à être la tête lexicale de plusieurs constituants. Dans ce cas, lors du passage à un arbre de dépendance, on écrase les différents constituants qui possèdent la même tête et la structure de dépendance est donc structurellement plus pauvre, bien qu'on puisse récupérer cette information (l'appartenance à une projection de la tête et pas à une autre) autrement, par exemple dans l'étiquetage des dépendances, en considérant différents types de relations syntaxiques comme cela est l'usage en grammaire de dépendance.

> La *tête syntaxique* d'un constituant est l'élément qui détermine la *valence passive* de ce constituant, c'est-à-dire l'élément qui contrôle la distribution de ce constituant.

Nous allons illustrer cette définition par des exemples. Commence par la tête de la phrase. A l'intérieur de la proposition, tout le monde s'accorde à considérer que le verbe fini est la tête, car c'est bien la présence du verbe fini qui fait qu'il s'agit d'une proposition. Signalons néanmoins deux difficultés :

1) Lorsque le verbe est composé (comme dans *Pierre **a donné** un livre à Marie*), on peut s'interroger sur qui de l'auxiliaire ou du participe est la tête de la proposition. Certains considèrent que l'auxiliaire dépend du participe. Je serais plutôt enclin à préférer considérer, à la suite de Tesnière ou de Mel'čuk, que l'auxiliaire est la tête. En effet, c'est l'auxiliaire qui porte le mode (*Il faudrait que Pierre ait/*a donné …*), qui hérite de marques grammaticales (*Pierre pense av**oir** donné …* ; *Pierre a**-t-il** donné…?*), qui porte la négation (*Pierre **n'a pas** donné…*), qui peut rester seul (*Pierre a-t-il donné …? Oui, il **a**.*), …

2) Dans des langues comme le français ou l'anglais où la présence du sujet est obligatoire, des linguistes ont été amenés à considérer que la présence du sujet n'était pas légitimé par le verbe, mais par un principe supérieur. Dans la grammaire générative, on considère à l'heure actuelle que le sujet, à la différence des compléments n'est pas gouverné par le verbe en tant que tel, mais par le morphème grammatical exprimant le temps. Dans la mesure où ce morphème appartient à la forme verbale et où en grammaire de dépendance on ne considère que les dépendances entre mots, les deux approches sont compatibles. Elles le sont encore en considérant que lorsque le verbe est composé, le sujet dépend de l'auxiliaire qui est aussi le porteur de la flexion temporelle.

Plus généralement, tout le monde s'accorde sur le sens de la relation de dépendance lorsqu'il existe une relation de *subordination*, c'est-à-dire lorsqu'il existe une relation actancielle (entre une tête et son *actant*) ou une relation modificative (entre une tête et un *modifieur*) (même si la frontière entre actant et modifieur est parfois difficile à saisir).[5] Pose problème la coordination et les relations avec des éléments jouant un rôle grammatical, notamment les complémenteurs, les déterminants et les auxiliaires. Avant de parler de la coordination, nous allons aborder la question des éléments grammaticaux en évoquant la théorie de la translation de Tesnière (1959).

Si l'œuvre de Tesnière est bien connue pour ce qui concerne la dépendance, on a souvent oublié sa théorie de la translation qu'il considérait probablement comme sa découverte principale (bien qu'on puisse estimer que l'idée est déjà là dans la théorie des rangs de Jespersen 1924). Selon Tesnière, il existe 4 parties du discours majeures (verbe, nom, adjectif, adverbe) avec des relations prototypiques entre ces parties du discours : les actants du verbe sont des noms et ses modifieurs des adverbes, les dépendants du nom sont des adjectifs et les dépendants de l'adjectif et de l'adverbe sont des adverbes. Néanmoins, un élément de partie du discours X peut venir occuper une position normalement réservée à un élément de partie du discours Y, mais dans ce cas, l'élément doit être *translaté* de la partie du discours X à la partie du discours Y par un élément morphologique ou analytique appelé un *translatif* de X en Y. Comme il y a 4 parties du discours majeures, il y aura 16 types de translatifs (y compris des translatifs de X en X qui ne change pas la partie du discours). Par exemple un verbe peut être l'actant d'un autre verbe (c'est-à-dire occuper une position nominale), mais il devra être à l'infinitif ou être accompagné de la conjonction de subordination *que* (*Pierre veut **la parole*** ; *Pierre veut **parler*** ; *Pierre veut **que** Marie **parle***). L'infinitif et la conjonction de subordination *que* sont donc des translatifs de verbe en nom. De même, les participes passé et présent, qui permettent à un verbe de modifier un nom (*le livre **rouge*** ; *le livre **volé** par Pierre* ; *la personne **volant** le livre*), sont

---

[5] La distinction entre actants et modifieurs est considérée par Tesnière (1959), à qui l'on doit le terme d'actant (et de *circonstant* pour les modifieurs). Nous reviendrons dans la Section 3.2 sur cette distinction qui joue un grand rôle dans la théorie Sens-Texte.

des translatifs de verbe en adjectif, la copule étant à son tour un translatif d'adjectif en verbe (*le livre **est** rouge* ; *le livre **est** volé par Pierre*). Les prépositions quant à elles seront catégorisées comme des translatifs de nom en adjectif ou en adverbe *(le livre **rouge*** ; *le livre **de Pierre*** ; *Pierre boit **maladroitement*** ; *Pierre boit **avec maladresse***).

Les cas de translation suscitent généralement des discussions quant au choix de la tête : le translatif, lorsqu'il est analytique doit-il être traité comme le gouverneur du translaté ou comme un dépendant ? Si l'on s'en tient à notre définition de la tête, le translatif doit être clairement considéré comme le gouverneur, car c'est bien lui qui contrôle la valence passive, son rôle étant justement de permettre au translaté d'occuper des positions auxquelles il ne pourrait accéder sans être translaté. Néanmoins, certains, comme Pollard & Sag (1994:44), considère que la conjonction de subordination *que* doit être traitée comme un marqueur, le verbe restant la tête de la complétive, arguant du fait que la distribution de la complétive dépend également du mode qui est porté par le verbe (*Il faut **que** Pierre **parte**/*part* ; *Marie pense **que** Pierre **part**/*parte*). En fait, cela revient à traiter les deux éléments, le translatif et le translaté, plus ou moins comme des co-têtes, puisque les traits tête et marqueur sont tous les deux des traits de tête (c'est-à-dire des traits dont les valeurs montent sur la structure résultante de leur combinaison). Tesnière lui-même hésite à traiter le translatif comme le gouverneur du translaté et préfère parler de *nucléus translatif* : il représente alors le translatif et le translaté comme un groupe (dessiné horizontalement) et dépendant ensemble de leur gouverneur. En plus, du fait que le translaté contrôle aussi quelque peu la distribution du groupe translatif-translaté (par exemple, certaines positions n'acceptent que des verbes infinitifs, c'est-à-dire des verbes translatés en nom, mais pas de noms : *Pierre peut **partir*** ; **Pierre peut le départ*), Tesnière argue du fait que les translatifs ont tendance à être analytique au départ et à se morphologiser par la suite (c'est-à-dire à devenir des morphèmes flexionnels sur le mot qu'il translate)[6] et que le lien entre le translatif et le translaté est particulièrement étroit.

La coordination est un autre cas qui pose problème. Si l'on s'en tient à considérer que la structure syntaxique doit être un arbre de dépendance et que tout groupe doit avoir une tête, le meilleur candidat est sans conteste le premier conjoint (Mel'čuk 1988a). Certains proposent également de prendre la conjonction de coordination comme tête du groupe coordonné, mais il s'agit alors d'un choix davantage guidé par la sémantique, la conjonction de coordination agissant comme opérateur sémantique prenant les conjoints comme arguments.[7] Mais on peut aussi considérer comme le font la plupart des grammaires syntagmatiques avec tête (Jackendoff 1977, Pollard & Sag 1994) que les conjoints sont des co-têtes. C'est également ce que propose Tesnière, bien que sa solution reste très informelle. Cf. Kahane 1997 pour voir comme la notion d'arbre de dépendance peut-être étendue pour prendre en compte cette hypothèse sans renoncer au traitement par un arbre de dépendance dans les autres cas. Parmi les autres cas qui posent problème citons le cas du déterminant (Zwicky 1985 ; Abney 1987 où il est défendu que le groupe nominal a le déterminant pour tête) et du pronom relatif (Tesnière 1959:561, Kahane & Mel'čuk 1999, Kahane 2000a).

Si comme on l'a vu, le recours à un arbre de dépendance peut dans certains cas ne pas être entièrement satisfaisant, je voudrais insister sur le fait que même dans ces cas-là, l'arbre de dépendance reste un moyen d'encodage suffisant. Il est possible que des moyens d'encodage de

---

[6] Tesnière distingue les translatés des dérivés : le translaté, même lorsque la translation est morphologique, continue à se comporter vis-à-vis de ses dépendants comme un élément de sa partie du discours initiale : par exemple, le verbe à l'infinitif, c'est-à-dire le verbe translaté en nom, continue à se comporter comme un verbe vis-à-vis de ses dépendants, à la différence du dérivé (*voler **un livre** est répréhensible* ; *le vol **de livre** est répréhensible*).

[7] De la même façon, un adjectif agit comme un prédicat sémantique qui prend le nom qu'il modifie comme argument (*le livre rouge* : rouge(livre)) sans qu'on souhaite pour autant considérer l'adjectif comme le gouverneur syntaxique du nom.

l'organisation syntaxique plus puissants permettent des analyses plus élégantes, mais l'arbre de dépendance conserve l'avantage de la simplicité. En plus, l'arbre de dépendance n'a pas, à la différence du rôle donné à l'arbre syntagmatique en grammaire générative, comme objectif d'encoder toutes les informations pertinentes sur une phrase. Dans la plupart des grammaires de dépendance et en particulier dans la théorie Sens-Texte, l'arbre de dépendance est avant tout une représentation intermédiaire entre la représentation sémantique et la représentation morphologique (là où les mots sont formés et ordonnés). L'arbre de dépendance doit donc contenir suffisamment d'information pour exprimer la relation avec la représentation sémantique, notamment les possibilités de redistribution ou de pronominalisation. De l'autre coté, il doit également contenir suffisamment d'informations pour exprimer les relations avec la représentation morphologique, c'est-à-dire les différentes possibilités d'ordre, d'accord ou d'assignation de cas. Ni plus, ni moins.

2.2 Fonctions syntaxiques

Un arbre de dépendance ne suffit pas à encoder l'organisation syntaxique des phrases sans un étiquetage des dépendances par des fonctions syntaxiques. Les *fonctions syntaxiques* permettent de distinguer les différents dépendants d'un même mot, mais aussi de rapprocher deux dépendants de deux mots différents qui présentent des comportements similaires vis-à-vis de différentes propriétés syntaxiques : placement, pronominalisation, régime, accord, redistribution, cooccurrence, … La notion de fonction syntaxique a été élaborée et utilisée indépendamment des grammaires de dépendance (cf. , par exemple, Jespersen 1924), même si la théorie de Tesnière a certainement marqué une étape fondamentale dans la compréhension de cette notion. Dans le courant génératiste, on a évité le recours explicite à un étiquetage fonctionnel en tentant d'encoder les différences de comportement d'un dépendant d'un mot par des différences de position dans l'arbre syntagmatique (par exemple le sujet est le GN sous S ou Infl', alors que l'objet direct est le GN sous GV). Néanmoins, de nombreuses théories issues de la grammaire syntagmatique (notamment LFG et HPSG) ont réintroduit explicitement la notion de fonction syntaxique, notamment à la suite des travaux de Comrie & Keenan 1987 sur la hiérarchie fonctionnelle.[8] (Cf. Abeillé 1996-97 pour un survol des différents arguments pour l'usage des fonctions syntaxiques en grammaire syntagmatique.)

L'une des principales difficultés pour décider combien de fonctions syntaxiques il est nécessaire de considérer est qu'on peut toujours attribuer une propriété particulière à la catégorie du dépendant ou du gouverneur (comme le font les grammaires syntagmatiques) plutôt qu'à l'étiquette de la relation de dépendance entre eux. Quitte à multiplier les catégories syntaxiques, il est formellement possible de limiter l'étiquetage des relations à un simple numérotage (il faut quand même garder un minimum pour distinguer entre eux les différents compléments du verbe). Il semble donc difficile d'établir des critères exacts pour décider si deux dépendances doivent ou non correspondre à la même fonction et il est nécessaire de prendre en compte l'économie générale du système en cherchant à limiter à la fois le nombre de catégories syntaxiques et le nombre de fonctions syntaxiques et à chercher la plus grande simplicité dans les règles grammaticales. On attribuera donc à la catégorie syntaxique les propriétés intrinsèques d'une lexie (c'est-à-dire qui ne dépendent pas de la position syntaxique) et à la fonction les propriétés intrinsèques d'une position syntaxique (c'est-à-dire qui ne dépendent pas de la lexie qui l'occupe). Autrement dit, on attribuera la même catégorie à des lexies qui présentent un comportement similaire dans toutes les positions syntaxiques et la même fonction à des positions syntaxiques qui présentent des comportements similaires avec toutes les lexies.

Pour caractériser l'ensemble des différentes fonctions syntaxiques, nous avons besoin de critères pour décider 1) si deux dépendants d'un même mot (dans deux phrases différentes)

---

[8] Toute grammaire syntagmatique qui fait usage des fonctions syntaxiques définit un arbre à la Gladkij (chaque syntagme dépend d'un mot) et devient de fait une grammaire de dépendance.

remplissent la même fonction et 2) si deux dépendants de deux mots différents remplissent la même fonction.

Pour le premier cas considérons le paradigme suivant : *Pierre lit* ***le livre*** / *Pierre* ***le*** *lit* / *le livre* ***que*** *Pierre lit.* On admet généralement que les syntagmes *un livre*, *le* et *que*[9] sont des réalisations du deuxième argument sémantique du prédicat *lire* ; plus précisément, *le* et *que* sont des formes pronominales de cet argument. De plus, ces syntagmes s'excluent mutuellement (**ce que Pierre lit le livre* ; **ce que Pierre le lit* ; seul *Pierre le lit le livre* est possible, mais avec une prosodie sur *le livre* très différente de *Pierre lit le livre*, qui laisse à penser que *le livre* ne remplit pas alors la même fonction). Dans ce cas, on considère que ces éléments remplissent tous la même fonction (à savoir la fonction d'*objet direct*).

Pourtant, les compléments *un livre*, *le* et *que* ne se positionnent pas de la même façon et les pronoms, à la différence des groupes nominaux, distinguent les cas (*il*/*le*/*lui* ; *qui*/*que*). Peut-être, peut-on distinguer fonction et relation syntaxique et dire que le clitique *le* remplit la fonction d'objet direct, mais dépend de son gouverneur par une relation spécifique (comme *objet-clitique*) qui impose un placement particulier ainsi que l'assignation d'un cas.[10] Dans ce cas, l'arbre de dépendance sera étiqueté par des relations syntaxiques et deux éléments remplissant des fonctions syntaxiques similaires pourront dépendre de leur gouverneur par des relations syntaxiques différentes. Néanmoins, il ne semble pas nécessaire de leur attribuer des relations syntaxiques différentes, car *le* comme *que* appartiennent à des classes fermées de mots outils pour lesquels on peut donner facilement des règles d'ordre spécifique. Mais on peut comprendre que certains préfèrent introduire des relations syntaxiques spécifiques pour ces éléments plutôt que de devoir invoquer des propriétés catégorielles de l'élément dépendant dans la règle de placement de l'objet direct.

Considérons un deuxième paradigme : *Pierre veut* ***un bonbon*** / *Pierre veut* ***manger*** / *Pierre veut* ***qu'on lui donne un bonbon***. Encore une fois, ces différents compléments réalisent tous le deuxième argument sémantique du verbe *vouloir*, s'excluent mutuellement et se pronominalisent de la même façon (*Pierre* ***le*** *veut* ; ***Que*** *veut Pierre ?*), ce qui nous inciterait à leur attribuer la même fonction syntaxique. Néanmoins, la construction avec verbe infinitif (*veut manger*) nécessite des spécifications supplémentaires, à savoir qu'il s'agit d'une construction à verbe contrôle, ou equi-construction, où le sujet du verbe *vouloir* coïncide avec le "sujet" de l'infinitif. Ceci peut suffire à vouloir introduire une relation particulière, bien qu'il existe d'autres façons d'encoder cette propriété (par exemple, en considérant directement une relation particulière entre le verbe infinitif et le sujet de *vouloir*).

Notons qu'il existe un autre critère souvent invoqué pour décider si deux dépendants d'un même mot remplissent la même fonction : la coordination (Sag *et al.* 1985, Hudson 1988). On peut décider par exemple que deux syntagmes peuvent être coordonnés seulement s'ils remplissent la même fonction (condition à laquelle s'ajouteront d'autres conditions, notamment sur l'identité catégorielle).[11] Dans notre dernier exemple, le fait que la coordination soit possible (*Pierre veut* ***un bonbon et manger***) nous incitera encore davantage à utiliser la même fonction.

---

[9] On suppose ici que *que* est traité comme un dépendant du verbe, ce qui n'est pas nécessairement justifié (Kahane 2000b).

[10] Pour assurer la montée du clitique dans, par exemple, *Pierre le fait lire à Marie*, on peut même considérer que le clitique dépend de *faire*, alors qu'il remplit une fonction vis-à-vis de *lire*.

[11] Par exemple, la coordination des adjectifs épithètes obéit à des conditions complexes et l'itération de la relation d'épithète est souvent préférable à la coordination : *des plats français exquis*, *?*des plats français et exquis*, *des plats français et néanmoins exquis*.

Considérons maintenant le deuxième cas : comment décider si des dépendants de deux mots différents doivent recevoir la même fonction. On considère que les dépendants de deux mots différents remplissent la même fonction si et seulement si ils acceptent les mêmes redistributions, les mêmes pronominalisations et les mêmes linéarisations (Iordanskaja & Mel'čuk 2000).

Considérons un premier exemple : *Pierre compte* ***sur Marie*** / *Pierre pose le livre* ***sur la table*** / *le livre est* ***sur la table***. Les dépendants *sur Marie* et *sur la table* remplissent-ils la même fonction ? Ces dépendants se distinguent nettement par leurs possibilités de pronominalisation : seul le deuxième accepte la cliticisation en *y* (**Pierre y compte ; Pierre y pose le livre* ; *le livre y est*) et les interrogatives et les relatives en *où (*Où Pierre compte-t-il ?* ; *Où Pierre pose-t-il le livre ?* ; *Où le livre est-il ?*). On distinguera donc deux fonctions syntaxiques différentes, *complément oblique* pour *compter* et *complément locatif* pour *poser* et *être* (qui n'est pas le même *être* que la copule).

Deuxième exemple : *Pierre compte* ***sur Marie*** / *Pierre est aidé* ***par Marie***. Les dépendants *sur Marie* et *par Marie* remplissent-ils la même fonction ? Aucune redistribution de ces dépendants n'est possible. On peut objecter que *Pierre est aidé par Marie* est le résultat de la passivation de *Marie aide Pierre*, mais la passivation est en quelque sorte orientée et *Marie aide Pierre* n'est pas le résultat d'une redistribution de *Pierre est aidé par Marie*. Les possibilités de pronominalisation sont les mêmes : pas de cliticisation, même pronominalisation pour les interrogatives et les relatives. On pourrait objecter que *sur N* accepte la pronominalisation en *dessus*, mais celle-ci est très régulière et doit être imputée à la préposition *sur* (de même qu'on aura *dessous* pour *sous* ou *dedans* pour *dans*) plutôt qu'à la fonction de *sur N*. Les possibilités de placement sont également les mêmes. En conséquence on peut attribuer à ces deux relations la même étiquette, par exemple *complément oblique*. Cela n'empêche pas de dire que *par Marie* dans *Pierre est aidé par Marie* est un complément d'agent ; cela ne signifie pas que ce groupe remplit la fonction syntaxique de complément d'agent qui n'a pas de raison d'exister en tant que telle, mais simplement que ce groupe est le résultat d'une réalisation particulière de l'"agent" suite à une redistribution.

Un dernier exemple : *Pierre mange* ***un bonbon*** / *Pierre veut* ***un bonbon***. Les deux dépendants *un bonbon* remplissent-ils la même fonction ? Les deux dépendants partagent les mêmes propriétés à une exception près, la passivation (*le bonbon est mangé par Pierre* ; ?**le bonbon est voulu par Pierre*). Deux solutions sont alors possibles : 1) considérer qu'il s'agit de la même fonction dans les deux cas (*objet direct*) et faire assumer la différence à la catégorie du verbe qui gouverne cette position ou 2) considérer qu'il s'agit de deux fonctions différentes. Etant donnée la grande similitude comportement, la première solution est plus économique.

En conclusion, comme on l'a vu, le choix d'un ensemble de fonctions syntaxiques est directement lié à la façon dont seront écrites les règles de pronominalisation, linéarisation, redistribution ou coordination.

2.3 Premières grammaires de dépendance

Dans cette section, nous allons présenter les premières grammaires de dépendance (Hays 1960, Gaifman 1965), qui ont pour particularité de ne traiter que des structures projectives.

Rappelons que l'un des points remarquables de la théorie de Tesnière est d'avoir dissocié la représentation syntaxique de l'ordre linéaire des mots : les arbres de dépendance de Tesnière ne sont pas ordonnés. L'objet de la syntaxe est alors d'exprimer le lien entre l'ordre des mots et leurs relations de dépendance.

L'une des principales propriétés de compatibilité entre un arbre de dépendance et un ordre linéaire est la *projectivité* (Lecerf 1961, Iordanskaja 1963, Gladkij 1966). Un arbre de dépendance assorti d'un ordre linéaire sur les nœuds est dit *projectif* si et seulement si, en

plaçant les nœuds sur une ligne droite et tous les arcs dans le même demi-plan, on peut assurer que 1) deux arcs ne se coupent jamais et que 2) aucun arc ne couvre la racine de l'arbre (Figure 4)[12].

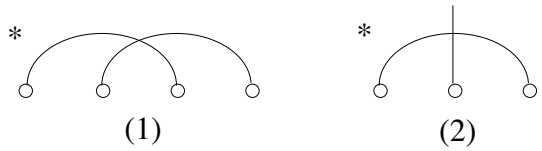

Figure 3 : Les cas de non projectivité

La projectivité est équivalente au fait que la *projection* de tout nœud x de l'arbre (c'est-à-dire l'ensemble des nœuds dominés par x, x compris) forme un segment continu de la phrase (Lecerf 1961, Gladkij 1966). Autrement dit, la projectivité dans le cadre des grammaires de dépendance correspond à la continuité des constituants dans le cadre des grammaires syntagmatiques. La littérature sur les structures de dépendance non projectives est d'ailleurs toute aussi abondante que la littérature sur les constituants discontinus (toutes proportions gardées). Nous y reviendrons à la fin de cette section.

La projectivité présente un intérêt immédiat : il suffit, pour ordonner un arbre projectif, de spécifier la position de chaque nœud par rapport à son gouverneur, ainsi que vis-à-vis de ses frères (Figure 4). Nous allons voir comment cette propriété est exploitée par les premières grammaires de dépendance.

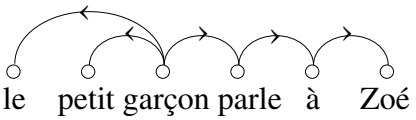

Figure 4 : Un exemple d'arbre de dépendance projectif

La première grammaire de dépendance formelle est due à Hays (1960). Une grammaire de Hays est constituée d'un vocabulaire V, d'une ensemble de catégories lexicales C, d'un lexique associant à chaque élément du vocabulaire une catégorie et d'un ensemble de règles de la forme $X(Y_1Y_2\dots Y_k*Y_{k+1}\dots Y_n)$ où X et les Y_i sont des catégories lexicales. La règle $X(Y_1Y_2\dots Y_k*Y_{k+1}\dots Y_n)$ indique qu'un nœud de catégorie X peut posséder n dépendants de catégories respectives Y_1, Y_2, …, Y_n placés dans l'ordre linéaire $Y_1Y_2\dots Y_k*Y_{k+1}\dots Y_n$ (où * indique la place de X par rapport à ses dépendants). Une règle de la forme X(*) indique qu'un nœud de catégorie X peut être une feuille de l'arbre de dépendance. Une telle grammaire permet de générer des arbres de dépendance projectifs dont les nœuds sont étiquetés par un mot de V et sa catégorie syntaxique dans C ou, ce qui revient au même, à générer des suites de mots de V où chaque mot correspond à un nœud d'un arbre de dépendance étiqueté par une catégorie syntaxique dans C. Comme on le voit les grammaires de Hays n'ont pas recours aux fonctions syntaxiques et elles génèrent simultanément des arbres de dépendances et des suites de mots. Comme l'a remarqué Gaifman (1965), les grammaires de Hays peuvent être simulées par des grammaires catégorielles à la Ajdukiewicz-Bar-Hillel (Ajdukiewicz 1935 ; Bar-Hillel 1953), la règle $X(Y_1Y_2\dots Y_k*Y_{k+1}\dots Y_n)$ correspondant simplement à la catégorie complexe $Y_k\dots Y_1\backslash X/Y_n\dots Y_{k+1}$ (l'inversion dans l'ordre des catégories est due au fait que la catégorie la plus à l'extérieur sera la première a être réduite et donnera donc le dépendant le plus proche de X). Si les grammaires catégorielles à la Ajdukiewicz-Bar-Hillel ne sont pas considérées comme

---

[12] Suivant Hudson 2000, nous représenterons la racine de l'arbre avec une dépendance gouverneur verticale (potentiellement infinie). La condition (2) se ramène alors à un cas particulier de la condition (1).

les premières grammaires de dépendance, c'est que les auteurs n'ont jamais mis leur formalisme en relation avec la construction d'arbres de dépendance (ni d'arbres syntagmatiques d'ailleurs). De plus, une catégorie complexe comme la catégorie N/N donnée à un adjectif antéposé ne s'interprète pas par " un adjectif est un N dont dépend un N à droite", mais comme "un adjectif est un mot qui combiné à un N à sa droite donne un syntagme de même nature" (voir néanmoins Lecomte 1992 pour une interprétation des grammaires catégorielles en termes de graphes et Rétoré 1996 pour le lien entre grammaire logique et réseaux de preuve, eux-mêmes interprétables en termes de graphes de dépendance). Gaifman (1965) a également noté que les grammaires de Hays sont trivialement simulables par des grammaires de réécriture hors-contextes où la règle $X(Y_1Y_2\ldots Y_k{*}Y_{k+1}\ldots Y_n)$ correspond à un famille de règle de réécriture $X\rightarrow Y_1Y_2\ldots Y_kaY_{k+1}\ldots Y_n$ pour tout mot a de catégorie X. Les grammaires de Hays, les grammaires d'Ajdukiewicz-Bar-Hillel et leurs équivalents en grammaire de réécriture se distinguent par la façon dont le vocabulaire pointe sur les règles syntaxiques.

L'article de Gaifman (1965) contient également deux résultats remarquables : l'équivalence faible entre les grammaires hors-contexte et les grammaires de dépendance de Hays[13] et un théorème d'équivalence forte entre une large classe de grammaires hors-contextes et les grammaires de dépendance de Hays (cf. également Dikovsky & Modina 2000).

D'un point de vue linguistique, les grammaires de Hays présentent plusieurs faiblesses : elles ne séparent pas les règles de bonne formation des arbres de dépendance des règles de linéarisation, c'est-à-dire des règles de mise en correspondance d'un arbre de dépendance et d'un ordre linéaire. De plus, concernant la bonne formation des arbres de dépendance, elles ne distinguent pas la sous-catégorisation et la modification. Ceci peut être résolu très simplement en divisant une règle de la forme $X(Y_1Y_2\ldots Y_k{*}Y_{k+1}\ldots Y_n)$ en trois familles de règles : une règle indiquant quels sont les catégories des s de X, des règles indiquant quels sont les catégories des modifieurs potentiels de X et une ou des règles indiquant comment les dépendants de X se placent les uns par rapport aux autres. On aura alors avantage à étiqueter les dépendances par des fonctions et à mentionner les fonctions plutôt que les catégories dans les règles de linéarisation. Nous verrons dans la suite comment ces différentes règles se présentent dans le cadre de la théorie Sens-Texte.

Enfin, les grammaires de Hays ne prévoient pas le traitement de structures non projectives. Pour traiter les cas non projectifs, différentes extensions sont possibles : on peut introduire des traits Slash dans les catégories comme cela est fait en GPSG et HPSG (Pollard & Sag 1994 ; cf. Lombardo & Lesmo 2000 pour une adaptation du procédé aux grammaires de dépendance), proposer des règles spécifiques qui permettent de déplacer des éléments dans l'arbre de dépendance pour se ramener à un arbre projectif (Hudson 2000, Kahane *et al.* 1998) ou utiliser une structure plus complexe où est vérifié un équivalent de la projectivité (Kahane 2000a). D'autres méthodes consistent à ne pas mettre en relation l'arbre de dépendance directement en relation avec l'ordre linéaire, mais à utiliser une structure syntagmatique intermédiaire comme cela est fait en LFG (Bresnan 1982, Bresnan *et al.* 1982 ; cf. Gerdes & Kahane 2001 ou Duchier & Debusman 2001 pour des méthodes équivalentes dans le cadre des grammaires de dépendance).

---

[13] Plus précisément, Gaifman 1965 montre que toute grammaire hors contexte est simulable par une grammaire dont les règles sont de la forme $X\rightarrow aY_1Y_2$, $X\rightarrow aY_1$ et $X\rightarrow a$, ce qui est un théorème bien connu sous le nom de théorème de mise en forme normale de Greibach, théorème attribué à Greibach (1965) par qui ce résultat a été démontré indépendamment.

3 Présentation de la théorie Sens-Texte

La théorie Sens-Texte [TST] est née il y a 35 ans des premiers travaux en traduction automatique en URSS (Žolkovskij & Mel'čuk 1965, 1967) et s'est depuis développée autour d'Igor Mel'čuk (Mel'čuk 1974, 1988a, 1997). Cf. également, pour d'autres présentations, Milićević 2001 ou Weiss 1999. La TST est intéressante à étudier dans le cadre d'une présentation des grammaires de dépendance, non seulement parce qu'il s'agit d'une des théories majeures utilisant des arbres de dépendance comme représentations syntaxiques, mais parce que les postulats même de la théorie conduisent naturellement à considérer une telle structure, où le mot joue un rôle central et où la structure syntaxique doit rendre compte des relations entre les mots. L'approche de la TST se distingue des grammaires syntagmatiques à plus d'un titre :

1) en privilégiant la sémantique sur la syntaxe ;
2) en privilégiant le sens de la synthèse sur celui de l'analyse pour la description ;
3) en donnant une grande importance au lexique (avec notamment la considération de la notion de fonction lexicale qui permet de décrire les relations lexicales dérivationnelles et collocationnelles) ;
4) en préférant une représentation syntaxique basée sur un arbre de dépendance plutôt qu'un arbre syntagmatique (ce qui est, en quelque sorte, une conséquence naturelle des points précédents).

Dans cette section, nous présenterons les postulats de base de la TST (Section 3.1), les différentes représentations d'une phrase considérées par la TST (Section 3.2) et les différentes règles d'un modèle Sens-Texte (Section 3.3). Dans la Section 4, nous présenterons une grammaire d'unification basée sur la TST.

3.1 Les postulats de base de la théorie Sens-Texte

La théorie Sens-Texte [TST] repose sur les trois postulats suivants.

Postulat 1. Une langue est (considérée comme) une correspondance multivoque[14] entre des sens et des textes[15].

Postulat 2. Une correspondance Sens-Texte est décrite par un système formel simulant l'activité linguistique d'un sujet parlant.

Postulat 3. La correspondance Sens-Texte est modulaire et présente au moins deux niveaux de représentation intermédiaires : le niveau syntaxique (structure des phrases) et le niveau morphologique (structure des mots).

Commentaires sur les postulats.

1) Le premier postulat de la TST signifie que la description d'une langue naturelle L consiste en la description de la correspondance entre l'ensemble des sens de L et l'ensemble des textes de L. On peut comparer ce point de vue à celui de Chomsky 1957, dont l'influence a été primordiale : la description d'une langue L consiste en un système formel dérivant l'ensemble des phrases (acceptables) de L. Pendant longtemps, ce point de vue a eu une interprétation plutôt restrictive,

---

[14] Plusieurs sens peuvent correspondre au même texte (homonymie) et plusieurs textes peuvent correspondre au même sens (synonymie).

[15] *Texte* renvoie à n'importe quel segment de parole, de n'importe quelle longueur, et *son* pourrait être un meilleur terme.

une phrase étant comprise comme une suite de caractères[16] — c'est-à-dire un texte dans la terminologie de la TST — ou au mieux comme une structure syntagmatique. Néanmoins, le postulat de Chomsky est formellement équivalent au premier postulat de la TST dès qu'on entend par phrase un signe au sens saussurien avec un signifié (le texte) et un signifiant (le sens). D'un point de vue mathématique, il est en effet équivalent de définir une correspondance entre l'ensemble des sens et l'ensemble des textes ou de définir l'ensemble des couples formés d'un sens et d'un texte en correspondance, un tel couple représentant une phrase[17] (Kahane 2000b, 2001).

2) Le deuxième postulat met l'accent sur le fait qu'une langue naturelle doit être décrite comme une correspondance. Un locuteur parle. Un modèle Sens-Texte (= le modèle d'une langue donnée dans le cadre de la TST) doit modéliser l'activité d'un locuteur, c'est-à-dire modéliser comment un locuteur transforme ce qu'il veut dire (un sens) en ce qu'il dit (un texte). C'est l'une des principales particularités de la TST de dire qu'une langue doit être décrite comme une correspondance (Sens-Texte) et, qui plus est, que la direction du sens au texte doit être privilégiée sur la direction du texte au sens.

3) Le troisième postulat de la TST appelle plusieurs commentaires. La plupart des théories linguistiques considèrent des niveaux de représentation syntaxique et morphologique. La particularité de la TST est de considérer que ces niveaux sont des niveaux *intermédiaires* entre le niveau sémantique (le sens) et le niveau phonologique (le texte). En conséquence, la correspondance entre les sens et les textes sera entièrement modulaire : une correspondance entre les niveaux sémantique et syntaxique, une correspondance entre les niveaux syntaxique et morphologique et une correspondance entre les niveaux morphologique et phonologique. (En fait, la TST considère non pas deux, mais cinq niveaux intermédiaires, ce qui ne change rien à notre discussion.)

Le résultat est que le module syntaxique, qui assure la correspondance entre les niveaux syntaxique et morphologique, ne fait qu'associer des représentations syntaxiques avec des représentations morphologiques. Il n'a pas pour objet, comme cela l'est pour une grammaire générative, de donner une caractérisation complète des représentations qu'il manipule. Dans le sens de la synthèse, le module syntaxique prend en entrée des représentations syntaxiques qui ont été synthétisées par le module sémantique à partir de représentations sémantiques bien formées et qui représentent des sens réels. En conséquence, une représentation syntaxique est caractérisée par l'ensemble des modules, par le fait qu'elle est un intermédiaire possible entre une représentation sémantique bien formée et une représentation phonologique correspondante. En conclusion, la TST ne donne aucune primauté à la syntaxe et la TST n'a pas pour objectif de donner une caractérisation explicite des représentations syntaxiques bien formées.

Je pense que, maintenant, 35 ans après leur première formulation, les postulats de la TST, même s'ils peuvent apparaître avec des formulations différentes, sont plus ou moins acceptés par l'ensemble de la communauté scientifique. Par exemple, j'aimerais citer les toutes premières phrase d'une monographie consacrée au Programme Minimaliste, la plus récente des théories chomskienne (Brody 1997) : “It is a truism that grammar relates sound and meaning. Theories that account for this relationship with reasonable success postulate representational levels corresponding to sound and meaning and assume that the relationship is mediated through complex representations that are composed of smaller units.” Le principal point qui semble ne

---

[16] Le meilleur exemple de cette interprétation restrictive du postulat de Chomsky est la définition du terme *langage formel* comme une suite de caractères. Un langage formel, pris dans ce sens, ne peut jamais modéliser l'essence d'une langue naturelle. En aucun cas, le fait de connaître l'ensemble des suites de caractères acceptables d'une langue ne peut être considéré comme la connaissance d'une langue ; il faut évidemment être capable d'associer ces suites à leur sens.

[17] Nous laissons de côté le fait que la description d'un langage ne se réduit pas à la description de phrases isolées.

pas être pris en considération par la plupart des descriptions formelles contemporaines des langues naturelles est le fait qu'une langue, si elle représente une correspondance entre des sens et des textes, doit être décrite par des règles de correspondance.

3.2 Niveaux de représentation

La TST sépare clairement les différents niveaux de représentation. Les représentations des différents niveaux ont des organisations structurelles différentes : les représentations sémantiques sont des graphes (de relations prédicat-argument), les représentations syntaxiques sont des arbres de dépendance (non ordonnés) et les représentations morphologiques sont des suites. Dans l'approche Sens-Texte, tout ce qui peut être différencié doit être différencié. Et des objets avec des organisations différentes doivent être représentés avec des moyens différents. De plus, la TST donne une grande importance à la géométrie des représentations. Le fait que les humains communiquent par la voix entraîne que les productions linguistiques sont irrémédiablement linéaires (même si à la suite des phonèmes se superpose la prosodie et si des gestes peuvent accompagner la parole). Par contre, tout laisse à penser que, dans notre cerveau tridimensionnel, le sens possède une structure multidimensionnelle. Le passage du sens au texte comprendrait alors, du point de vue de l'organisation structurelle, deux étapes essentielles : la hiérarchisation, c'est-à-dire le passage d'un sens multidimensionnel à une structure syntaxique hiérarchique (= bidimensionnelle), et la linéarisation, c'est-à-dire le passage de cette structure hiérarchique à une structure linéaire (= unidimensionnelle).

3.2.1 Représentation sémantique

Le sens est défini, dans le cadre de la TST, comme un invariant de paraphrase, c'est-à-dire comme ce qui est commun à toutes les phrases qui ont le même sens. Ceci fait automatiquement de la TST est un modèle de la paraphrase (Mel'čuk 1988b) et, par conséquent, un outil adapté à la traduction automatique (les deux sont intimement liées, la paraphrase étant de la traduction intralangue).

Le cœur de la *représentation*[18] *sémantique* est un graphe dont les nœuds sont étiquetés par des *sémantèmes*. Une représentation sémantique est un objet purement linguistique spécifique à une langue. Un *sémantème lexical* d'une langue L est le sens d'une lexie[19] de L dont le signifiant

---

[18] Le terme de *représentation*, utilisé par Mel'čuk lui-même, est en fait un peu contradictoire avec le point de vue de la TST. En un sens, la représentation sémantique ne représente pas le sens d'un texte, mais c'est plutôt les textes qui expriment des représentations sémantiques. Mel'čuk (2001:15) dit d'ailleurs à ce propos : "During the process of sentence construction (= synthesis), lexical and syntactic choices carried out by the Speaker very often lead to the modification of the starting meaning, i.e. of the initial semantic representation, making it more precise and specific: the lexical units bring with them additional nuances of meaning that have not been present in the initial semantic representation. The MTT tries to model this phenomenon; as a result, quite often the following situation obtains: Suppose that the synthesis starts with the representation 'σ' and produces sentences 'S_1', 'S_2', …, 'S_n'; the sentences having as their common source the semantic representation 'σ' are considered to be synonymous. Now if we analyze these sentences semantically, the semantic 'S_1', 'S_2', …, 'S_n' obtained from this process may well be different from each other and from the initial semantic representation 'σ' ! [...] The initial semantic representation is taken to be rather approximate—it need not necessarily fully specify the meaning of the sentences that can be obtained from it. The meaning can become more precise—or less precise—in the course of its lexicalization and syntacticization."

[19] Un vocable est ensemble de lexies correspondant aux différentes acceptions d'un même mot. En toute rigueur, le nom d'une lexie doit être accompagné, comme dans le dictionnaire, d'un numéro qui la distingue des autres lexies du vocable.

peut-être un mot ou une configuration de mots formant une locution. Par exemple, 'cheval', 'pomme de terre', 'prendre le taureau par les cornes' sont des sémantèmes du français. Des lexies de parties du discours différentes peuvent avoir le même sémantème ; ainsi, 'partir' = 'départ' ('j'attends ton départ' = 'j'attends que tu partes') ou 'durer' = 'pendant' ('Ta sieste a duré 2 heures' = 'Tu as fais la sieste pendant 2 heures')[20]. Il existe aussi des *sémantèmes grammaticaux* correspondant au sens des morphèmes flexionnels (ou de configurations contenant des morphèmes flexionnels, comme le passé composé) : par exemple, '*singulier*', '*défini*', '*présent*' ou '*passé composé*' sont des sémantèmes grammaticaux.[21]

Un sémantème agit comme un prédicat et est lié à ses arguments par des arcs pointant sur eux. Les différents arcs émergeant d'un sémantème sont numérotés de 1 à n, en suivant l'ordre d'oblicité croissant des arguments. Un arc représente une relation prédicat-argument et est appelée une *dépendance sémantique*. Les dépendances sémantiques doivent être distinguées des dépendances syntaxiques. Comme l'a noté Tesnière lui-même (1959:42), dans la plupart des cas, quand un mot B dépend syntaxiquement d'un mot A, il y a une dépendance sémantique entre 'A' et 'B'. Mais ce que n'avait pas vu Tesnière (et qui est probablement une découverte attribuable à Žolkovskij & Mel'čuk 1965), c'est que la dépendance sémantique peut être orientée de 'A' et 'B' comme de 'B' vers 'A'. Par exemple, dans *une petite rivière*, *petite* dépend syntaxiquement de *rivière*, mais, parce que la petitesse est une propriété de la rivière, 'rivière' est un argument du prédicat 'petit'. Par contre, dans *la rivière coule*, *rivière* dépend syntaxiquement de *coule* et, parce que l'écoulement est une propriété de la rivière, 'rivière' est un argument du prédicat 'couler'. Quand les dépendances sémantique et syntaxique sont dans la même direction, on dit que B est un *actant* de A (*rivière* est un actant de *coule* dans *la rivière coule*), tandis que, quand les dépendances sémantique et syntaxique sont dans la direction opposée, on dit que B est un *modifieur* de A (*petite* est un modifieur de *rivière* dans *une petite rivière*). Il existe aussi des cas où dépendances sémantique et syntaxique ne se correspondent pas, comme dans les phénomènes de montée (dans *Pierre semble malade*, *Pierre* dépend syntaxiquement de *semble*, mais 'sembler' est un prédicat unaire qui prend seulement 'malade' comme argument) ou de *tough*-movement (dans *un livre facile à lire*, *facile* dépend syntaxiquement de *livre*, mais 'livre' est un argument de 'lire' et pas de 'facile') ; voir également le cas des relatives et des interrogatives indirectes (Kahane & Mel'čuk 1999).

La *valence sémantique* d'un sémantème, c'est-à-dire l'ensemble de ses arguments sémantiques, est déterminée par sa définition lexicographique. Ainsi 'blessureI.2' est une prédicat ternaire (Mel'čuk *et al.* 1999 ; définition révisée) : 'blessureI.2 de X à Y par Z' = 'lésion à la partie Y du corps de X qui est causée par Z et qui peut causer une ouverture de la peau de Y, un saignement de Y, une douleur de X à Y ou la mort de X' (sa_X *blessure par balle*$_Z$ *à la jambe*$_Y$).

La représentation sémantique comprend, en plus du graphe sémantique, trois autres structures qui s'y superposent : la structure communicative, la structure référentielle (qui relie des portions du graphe aux référents qu'elles dénotent) et la structure rhétorique (qui indiquent les intentions stylistiques du locuteur, c'est-à-dire si celui-ci veut être neutre, ironique, relâché, humoristique, …). La Figure 5 présente une représentation sémantique simplifiée (limitée au graphe sémantique et à la thématicité) pour la phrase (1) :

(1) *Zoé essaye de parler à la belle dame*

---

[20] Les deux phrases peuvent apparaître non synonymes en raison de la structure communicative (voir plus loin) : par exemple, si la première peut facilement avoir pour thème 'la durée de ta sieste' (= 'ta sieste a duré'), cela paraît plus difficile pour la deuxième qui aura plutôt pour thème 'toi' ou 'ta sieste' (= 'tu as fait la sieste').

[21] Comme pour les lexies, les différentes acceptions d'un morphème flexionnel devraient être distinguées par des numéraux. A noter que Mel'čuk ne considère pas de sémantèmes grammaticaux et utilisent des paraphrases lexicales : 'plus d'un', 'avant maintenant', …

La même représentation sémantique vaut pour des paraphrases de (1) comme *Zoé cherche à dire un mot à la jolie femme*.

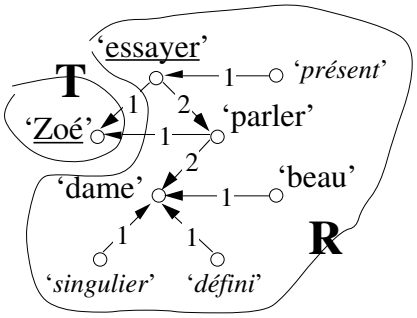

Figure 5 : La représentation sémantique de (1)

La *structure communicative* spécifie la façon dont le locuteur veut présenter l'information qu'il communique (de quoi il parle, ce qu'il veut dire, ce qu'il veut souligner, ce qu'il présente comme information commune avec son interlocuteur, …). La structure communicative est encodée en marquant certaines zones du graphe sémantique par des *marques communicatives*. Dans chacune de ces zones, on indique (par un soulignement) le sémantème qui résume le contenu sémantique de cette zone (Polguère 1990). Mel'čuk 2001 propose huit catégories communicatives : thématicité (thème-rhème-spécifieurs), donné-nouveau, focalisation, perspective (arrière-plan), emphatisation, présupposition, unitarité (unitaire-articulé) et locutionalité (signalé-performé-communiqué). Nous allons montrer comment des changements dans la thématisation et la focalisation du graphe sémantique de la Figure 5 donnent d'autres phrases. Tout message doit nécessairement communiquer quelque chose (le *rhème*) à propos de quelque chose (le *thème*) ou éventuellement de rien. La phrase (1) peut être glosée par 'à propos de Zoé (thème), je veux dire qu'elle essaye de parler à une belle dame (rhème)'. La partition thème-rhème s'identifie en voyant quelle est la question sous-jacente au message communiqué (ici 'que fait Zoé ?'). Un élément *focalisé* quant à lui est une partie du sens que le locuteur présente comme étant localement proéminente pour lui – ou, en d'autres termes, comme étant le siège (angl. *focus*) de son attention (Mel'čuk 2001). Voici quelques phrases ayant le même graphe sémantique que (1) avec des structures communicatives différentes (et pour lesquelles les mêmes choix lexicaux ont été faits) :

(3) a. *La belle dame, Zoé essaye de lui parler.* ('belle dame' thème focalisé)
b. *C'est à la belle dame que Zoé essaye de parler.* ('belle dame' rhème focalisé)
c. *Zoé, c'est à la belle dame qu'elle essaye de parler.* (focalisation de 'Zoé' en plus)
d. *Ce que Zoé essaye de faire, c'est de parler à la belle dame.* ('Zoé essaye' thème foc.)

On trouvera de nombreux exemples dans Mel'čuk 2001. Le rôle de la structure communicative dans la production des relatives est étudié dans Kahane & Mel'čuk 1999. Notons encore que la structure communicative est souvent considérée, à la différence de la structure prédicat-argument, comme très vague et difficile à cerner précisément. Nous pensons que cette vision est complètement fausse et due en partie au fait que les linguistes étudient généralement cette question du point de vue de l'analyse, en cherchant à déterminer la structure communicative de textes. Evidemment, si l'on considère un énoncé isolé tel que *Marie a dit que Pierre est parti*, il est impossible de déterminer sa partition thème-rhème. Cette phrase peut "répondre", avec certes des prosodies différentes, à des questions aussi diverses que *Que fait Marie ?, Qu'a dit Marie ?*, *Qu'a dit Marie de Pierre ?*, *Que fait Pierre ?*, *Qui a dit que Pierre est parti* ?, *Qui est parti ?*, … (par exemple, lorsqu'elle répond à la question *Que fait Pierre ?*, 'Pierre' est le thème, 'partir' est le rhème, 'Marie m'a dit' un *spécifieur* qui spécifie sous quelles conditions je peux dire que Pierre est parti). Par contre si on se place du point de vue de la synthèse, il est évident que le locuteur sait de quoi il veut parler et ce qu'il veut dire à ce propos. La partition thème-rhème est donc parfaitement établie (notons d'ailleurs qu'elle s'établit au niveau sémantique). Dû au

pouvoir paraphrastique de la langue, il reste au locuteur de nombreuses possibilités d'énonciation conditionnées aussi par les autres choix communicatifs, notamment la focalisation. Des énoncés tels que *Marie a dit que Pierre est parti*, peu conditionnés par la structure communicative, seront possibles avec de nombreux choix communicatifs, mais d'autres choix syntaxiques ne seront possibles qu'avec des choix communicatifs précis : *D'après Marie, Pierre est parti* (‘Marie a dit’spécifieur), *C'est Marie qui a dit que Pierre est parti* (‘Marie’ rhème focalisé), etc.

Une représentation sémantique peut être encodée dans un style inspiré de la logique. La traduction d'un graphe sémantique en une formule logique nécessite d'introduire une variable pour chaque nœud du graphe (à l'exception des nœuds étiquetés par un sémantème grammatical). Cette variable représente le nœud et est utilisée comme argument par tout sémantème pointant sur le nœud. En introduisant des variables x, y, p, e et e' pour les sémantèmes lexicaux ‘Zoé’, ‘dame’, ‘beau’, ‘essayer’ et ‘parler’, on peut encoder le graphe de la Figure 5 par la formule (2).

(2) THEME(x) RHEME(e)
x : ‘Zoé’ e : ‘essayer’(x,e')
e' : ‘parler’(x,y)
y : ‘dame’
p : ‘beau’(y)
‘*présent*’(e)
‘*singulier*’(y)
‘*défini*’(y)

La structure thème-rhème est encodée par la partition des sémantèmes en deux groupes et par les “prédicats” THEME et RHEME pointant sur les nœuds dominants de ces deux zones. Si l'on omet la structure thème-rhème, l'ordre des prédicats n'est pas pertinent et la formule s'apparente à une formule conjonctive du calcul des prédicats (cf. par exemple les représentations sémantiques de la DRT ; Kamp 1981, Kamp & Reyle 1993). La variable représentant un nœud peut d'ailleurs être attribuée au sémantème (on parle de réification) : ainsi à la place des notations y : ‘dame’ ou e : ‘essayer’(x,e'), on peut utiliser les notations ‘dame’(y) ou ‘essayer’(e,x,e'), plus habituelles en logique.

Malgré leur similitude formelle, les représentations sémantiques de la TST doivent être distinguées des représentations sémantiques des sémantiques issues de la logique frégéenne, comme la DRT. En TST, la représentation sémantique ne représente pas l'état du monde que dénote un sens, mais le sens lui-même. En particulier, les variables que nous avons introduites lors de la réification ne renvoient pas, comme c'est le cas dans la logique frégéenne, à des objets du monde. Les variables renvoient ici uniquement aux sémantèmes, c'est-à-dire aux signifiés des mots. Donnons un exemple : dans le sens de *une grosse fourmi*, le sémantème ‘gros’ est un prédicat unaire dont l'argument est le sémantème ‘fourmi’ et en aucun cas le référent de *fourmi*. D'ailleurs, quand on parle d'*une grosse fourmi*, on ne veut pas dire que le référent de fourmi est gros en soi (d'ailleurs rien n'est gros en soi), mais qu'il est gros en tant que fourmi. La chose est peut-être encore plus évidente quand on parle d'*un gros fumeur*. Ici non plus, on ne veut pas dire que le référent de *fumeur* est gros, mais que quelque chose dans le sens ‘fumeur’ est gros. En effet, si un ‘fumeur’ est une ‘personne qui fume (régulièrement)’, un ‘gros fumeur’ est une ‘personne qui fume (régulièrement) en grosse quantité’. D'autre part, le sémantème ‘gros’ pourra lui-même être l'argument d'un autre sémantème comme *dans une **très** grosse fourmi ou une fourmi **plus** grosse que mon pouce*, ce qui nécessite d'introduire une variable pour ‘gros’ lors de la réification, sans qu'on veuille pour autant considérer que *gros* possède un référent de discours.

En TST, le sens est défini comme ce qui est commun à tous les énoncés qui ont le *même sens*. La définition n'est pas circulaire, ‘avoir le même sens’ étant défini préalablement au ‘sens’ : il est plus facile de demander à un locuteur si deux énoncés ont le même sens (sont synonymes)

que de lui demander quel est le sens d'un énoncé (ce qui d'ailleurs le conduira essentiellement à proposer des énoncés qui ont le même sens). La TST est donc un modèle de la paraphrase. Le sens est un objet purement linguistique. La description du monde est reléguée à un niveau de représentation plus profond, extralinguistique. Remarquons tout de même que la référence au monde extérieur n'est pas exclue de la représentation sémantique de la TST et fait l'objet d'une structure particulière, la *structure référentielle*, superposée au graphe sémantique et indiquant quelle zone du graphe correspond à un référent de discours.

Notons enfin, pour terminer sur les différences entre les représentations sémantiques de la TST et les formules logiques, que tous les sémantèmes sont formalisés par des prédicats (les noms sémantiques comme 'Zoé' ou 'dame' étant des cas particuliers de prédicats à zéro argument), même des sens comme 'quel que soit', 'quelqu'un', 'et' ou 'non', qui sont habituellement formalisés en logique par des objets d'une autre nature, quantifieurs ou connecteurs.[22]

3.2.2 Représentation syntaxique profonde

Le niveau syntaxique profond est un niveau intermédiaire entre le niveau sémantique et le niveau syntaxique de surface, où le graphe a été hiérarchisé et les sémantèmes lexicalisés, mais où ne figure pas encore à proprement parlé les mots. Le cœur de la *représentation syntaxique profonde* est un arbre de dépendance (non ordonné) dont les nœuds sont étiquetés par des *lexies profondes* accompagnées chacune d'une liste de *grammèmes profonds*. Les *lexies profondes* sont des lexies pleines correspondant à des mots ou à des locutions. Les lexies vides, comme les prépositions régies, n'apparaissent qu'au niveau syntaxique de surface. De même, les *grammèmes profonds* sont des morphèmes grammaticaux pleins ; les grammèmes vides dus à l'accord ou à la rection, comme le cas, apparaissent plus tard. Les catégories grammaticales profondes du verbe sont le mode, le temps et la voix. Les lexies sont écrites en majuscules et les grammèmes placés en indice : $\text{LEXIE}_{\text{grammème}}$. Les branches de l'arbre sont étiquetées avec un petit ensemble de *relations syntaxiques profondes* : les actants sont simplement numérotés par oblicité croissante (I, II, III, …), les modifieurs reliés à leur gouverneur par la relation ATTR (angl. *attributive*) et deux autres relations sont considérées, COORD pour les groupes coordonnés (*Marie, Jean et Pierre* : MARIE $-\text{COORD}\rightarrow$ JEAN $-\text{COORD}\rightarrow$ ET $-\text{II}\rightarrow$ PIERRE) et APPEND pour les parenthétiques, les interjections, les interpellations, etc. (***Naturellement**, il n'a rien fait ; Où vas-tu, **Zoé** ?*). Kahane & Mel'čuk 1999 introduisent une autre relation pour les modifieurs qualitatifs (non restrictifs) et Kahane 1998 propose l'introduction d'une relation spécifique pour un actant rétrogradé (tel que le complément d'agent).

Nous proposons Figure 6 la représentation syntaxique profonde de (1) ; Le trait hachuré représente une relation de coréférence entre les deux occurrences de ZOÉ résultant de la coupure du graphe sémantique au niveau du sémantème 'Zoé'. L'une des deux occurrences sera effacée en surface par la règle de pronominalisation de l'actant I d'un verbe à l'infinitif. Le grammème infinitif sera introduit par le régime de ESSAYER au niveau syntaxique de surface. La structure

---

[22] La différenciation formelle des quantifieurs est certainement nécessaire pour la déduction logique, mais ne l'est pas forcément pour la paraphrase et la traduction. La portée des quantifieurs n'est pas clairement encodée dans les représentations sémantiques standard. En un sens, il n'est pas sûr que la portée des quantificateurs doive réellement être encodée dans la représentation sémantique de *Tous les hommes cherchent un chat* et il est curieux de voir fleurir des travaux qui montrent comment sous-spécifier les représentations sémantiques dans des formalismes qui obligent à indiquer la portée des quantifieurs. Mel'čuk 2001 émet l'hypothèse que les effets de portée des quantificateurs résultent de la structure communicative. Polguère 1992 propose d'encoder les quantifieurs comme des sémantèmes biactanciels dont le deuxième argument, représentant la portée, pointe sur une zone du graphe, ce qui pourrait être relié à l'hypothèse précédente. Dymetman & Coperman 1996 proposent une solution à l'encodage de la portée des quantifieurs avec une représentation intermédiaire entre graphe sémantique et formule logique.

communicative syntaxique profonde qui reprend la structure communicative sémantique n'est pas représentée ici.

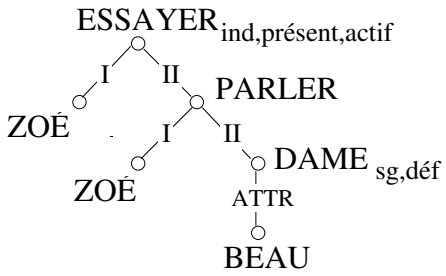

Figure 6 : Représentation syntaxique profonde de (1)

Notons encore l'une des spécificités de l'approche Sens-Texte : le concept de *fonction lexicale* (Žolkovskij & Mel'čuk 1965, Mel'čuk *et al.* 1995, Wanner 1996, Kahane & Polguère 2001). Certains sens, comme l'intensification, le commencement, la causation, la réalisation, etc., tendent à s'exprimer de manière collocationnelle, c'est-à-dire que leur expression n'est pas déterminée librement, mais dépend fortement de l'expression d'un de leurs arguments sémantiques. Les fonctions lexicales sont des "lexies" dont le signifiant n'est pas un mot précis, mais varie en fonction de l'expression d'un argument. Par exemple, le sens 'intense' pourra s'exprimer avec *amoureux* par *follement*, avec *heureux* par *comme un pape*, avec *improbable* par *hautement*, avec *blessé* par *gravement*, etc. De même, le sens 'commencer' pour s'exprimer avec *incendie* par *se déclarer*, avec *jour* ou *vent* par *se lever*, avec *orage* par *éclater*, etc. Les fonctions lexicales correspondantes seront notées **Magn** et **Incep**. Ces fonctions seront utilisées pour étiqueter un nœud de l'arbre syntaxique profond correspondant à un sens 'intense' ou 'commencer'. Les valeurs seront introduites seulement dans l'arbre syntaxique de surface (Mel'čuk 1988, Polguère 1998).

3.2.3 Représentation syntaxique de surface

Le cœur de la *représentation syntaxique de surface* d'une phrase est un arbre de dépendance (non ordonné) à la façon des arbres de dépendance de Tesnière 1959. Les nœuds de l'arbre sont étiquetés par des *lexies de surface* accompagnées chacune d'une liste de *grammèmes de surface*. Chaque lexie de surface correspond à un mot de la phrase. Ces lexies peuvent correspondre directement à une lexie profonde, ou bien correspondre à l'un des mots d'une locution, ou être la valeur d'une fonction lexicale, ou bien être une lexie vide introduite par un régime (comme les prépositions DE et À ici), ou encore être une partie de l'expression d'un grammème profond (comme un auxiliaire de temps ou l'article LE ici). Les grammèmes de surface correspondent directement à un grammème profond, sauf pour les grammèmes profonds qui ont une expression analytique comme les temps composés, les voix ou la détermination. Les grammèmes de surface d'accord ou de régime (comme les cas) ne sont introduits qu'au niveau morphologique profond. Les branches de l'arbre syntaxique de surface sont étiquetées par des fonctions syntaxiques ou *relations syntaxiques de surface* (cf. Mel'čuk 1974 pour le russe, Mel'čuk & Pertsov 1987 pour l'anglais et Iordanskaja & Mel'čuk 2000 pour le français; voir aussi Section 2.2).

Figure 7 : Représentation syntaxique de surface de (1)

3.2.4 Représentation morphologique profonde

Le cœur de la *représentation morphologique profonde* d'une phrase est la suite des représentations morphologiques des mots de la phrase, c'est-à-dire une *chaîne morphologique*. La représentation morphologique d'un mot est une lexie de surface accompagnée d'une liste de grammèmes de surface. A noter que, en français, l'adjectif s'accorde en genre et nombre avec le nom et le verbe en personne et nombre avec son sujet. Le nom possède une marque de genre qui est indiquée dans son entrée lexicale, mais ne porte pas de grammème de genre comme les adjectifs, car il n'est pas fléchi par le genre.[23]

La chaîne morphologique de (1) est :

(2) $ZOÉ\ ESSAYER_{ind,présent,3,sg}\ DE\ PARLER_{inf}\ À\ \ LE_{fém,sg}\ BEAU_{fém,sg}\ \ DAME_{sg}$

La représentation morphologique d'une phrase comprend, en plus de la chaîne morphologique, une structure prosodique. La *structure prosodique* au niveau morphologique est essentiellement un regroupement des mots en groupes prosodiques, agrémenté de marques prosodiques calculées en fonction des marques communicatives des portions de l'arbre syntaxique de surface auxquelles correspondent ces groupes. La véritable structure prosodique sera calculée au niveau phonologique à partir de cette structure prosodique de niveau morphologique et des propriétés phonologiques des mots (qui ne sont pas encore prises en compte au niveau morphologique, les phonèmes n'étant pas considérés). Dans Gerdes & Kahane 2001 (inspirés par des discussions avec Igor Mel'čuk), nous proposons de construire au niveau morphologique une structure syntagmatique qui, contrairement à l'usage qu'en font les grammaires basées sur la syntaxe X-barre, n'encode pas la représentation syntaxique de la phrase, mais plutôt sa structure prosodique de niveau morphologique (cf. aussi Section 6).

3.3 Modèle Sens-Texte standard

On appelle *modèle Sens-Texte* [MST] d'une langue le modèle de cette langue dans le cadre de la TST. Le terme *modèle* est préféré au terme plus couru de *grammaire* (*formelle*), car le terme *grammaire* masque le fait qu'une composante essentielle d'un modèle d'une langue, à côté de la grammaire proprement dite est le lexique.

---

[23] Il n'est probablement pas judicieux de considérer que les noms possèdent un trait de personne. On peut penser que seuls les pronoms possèdent un tel trait et que le verbe prend la 3ème personne par défaut, comme il le fait aussi avec les sujets verbaux ou phrastiques (*que tu viennes* ***est*** *une bonne surprise*).

Dans la Section 3.3.1, nous présenterons le lexique d'un MST, puis, dans la Section 3.3.2, l'architecture générale d'un module de correspondance. Enfin, dans les Sections 3.3.2 à 0, nous présenterons les modules sémantique, syntaxique profond et syntaxique de surface d'un MST.

3.3.1 Le lexique d'un modèle Sens-Texte

La TST donne une très grande importance au lexique. Le lexique d'un MST est appelé un *Dictionnaire Explicatif et Combinatoire* [DEC] . Un premier DEC pour le russe a été proposé par Mel'čuk & Žolkovsky 1984. Un DEC du français est maintenant en développement depuis 20 ans à l'université de Montréal (Mel'čuk *et al*. 1984, 1988, 1992, 1999). Les entrées du DEC sont les lexies profondes.[24] En plus de la description des valences sémantique et syntaxique, qui est indissociable des approches basées sur la dépendance, le DEC se caractérise par le grand soin donné à la définition sémantique des lexies (basée sur la paraphrase) et par l'utilisation des fonctions lexicales dans la description des liens dérivationnels et collocationnels entre lexies (Mel'čuk *et al.* 1995, Kahane & Polguère 2001).

Nous allons présenter et commenter l'entrée (révisée et simplifiée) de la lexie BLESSUREI.2 (Mel'čuk *et al*. 1999). Chaque article est divisé en trois zones :

- la *zone sémantique* donne la définition lexicographique de la lexie ;

- la *zone syntaxique* donne le tableau de régime (ou cadre de sous-catégorisation de la lexie, c'est-à-dire la correspondance entre les actants sémantiques (X, Y, …), les actants syntaxiques profonds (I, II, …) et les leur expression de surface (*à* N, *par* N, …); le tableau de régime est suivi par des conditions particulière (l'expression 1 de la colonne 3 n'est possible que si N est une arme blanche) et des exemples de réalisations et de combinaisons des différents actants ;

- la *zone de cooccurrence lexicale* donne les valeurs des fonctions lexicales pour la lexie, c'est-à-dire les collocations formées avec cette lexie (*une blessure cuisante, se faire une blessure, la blessure s'infecte*, …) et les dérivations sémantiques de la lexie (*blessé, plaie, se blesser*, …).

Exemple : article de dictionnaire de BLESSUREI.2

Définition lexicographique

'blessureI.2 de X à Y par Z' = 'lésion à la partie Y du corps de X qui est causée par Z et qui peut causer (I) une ouverture de la peau de Y, (II) un saignement de Y, (III) une douleur de X à Y ou (IV) la mort de X'[25]

---

[24] La description séparée des lexies de surface pourrait être également utile. Pour l'instant, celles-ci sont décrites grossièrement lorsqu'elle apparaissent dans la description des lexies profondes, comme élément d'une locution, comme valeur d'une fonction lexicale ou comme élément régi introduit dans le tableau de régime.

[25] La définition lexicographique est basée sur la paraphrase (la définition de L doit est substituable à L) et la cooccurrence lexicale : les composantes (I) à (IV) sont conditionnées par les différentes valeurs des Fact-Real de BLESSUREI.2. Cette portion de la définition indique les "objectifs" inhérents de L (*une blessure peut être profonde* (I), *saigner* (II), *faire souffrir* (III) *ou être fatale* (IV)). Voir, pour comparaison, les valeurs du trait *telic* dans les descriptions lexicales du Lexique Génératif de Pustejovsky 1995.

Régime

X = 1	Y = 2	Z = 3
1. *de* N 2. A_{pos}	1. *à* N	1. *à* N 2. *par* N

Contrainte sur 3.1 : N désigne une arme blanche
Contrainte sur 3.2 : N = *balle*, …

Exemples

1 : *la blessure de Jean/du soldat/du cheval* ; *sa blessure*
2 : *une blessure à l'épaule/au cœur/à l'abdomen* ; *des blessures au corps*
3 : *une blessure à l'arme blanche/au couteau* ; *une blessure par balle*
1 + 2 : *les blessures de l'enfant aux bras* ; *sa blessure au poignet droit*
1 + 2 +3 : *sa blessure par balle à la jambe*

Fonctions lexicales

$Syn_{\subset}$: lésion
$Syn_{\supset}$: coupure, écorchure; égratignure; morsure; brûlure; ecchymose; déchirure; fracture; entorse
$Syn_{\cap}$: plaie; bobo "fam"
personne-S_1 : blessé
$A_{1/2}$: // blessé
$A_{1/2}$+Magn : couvert, criblé [de ~s]
Magn : grave, majeure, sérieuse
AntiMagn : légère, mineure, superficielle // égratignure
AntiBon : mauvaise, vilaine
IncepMinusBon : s'aggraver; s'enflammer, s'envenimer, s'infecter
$Oper_1$: avoir [ART ~]; porter [ART ~]; souffrir [de ART ~]
$FinOper_1$: se remettre, se rétablir [de ART ~]
$Caus_1Oper_1$: se faire [ART ~]
$LiquOper_1$: guérir [N de ART ~]
$FinFunc_0$: se cicatriser, (se) guérir, se refermer
essayer de $LiquFunc_0$: soigner, traiter [ART ~]; bander, panser [ART ~]
$CausFunc_1$: faire [ART ~ à N]; infliger [ART ~ à N] // blesser [N] [avec N=Z]
$Caus_1Func_1$: se faire [ART ~]; se blesser [avec N=Z]
$Real_1$: (II) souffrir [de ART ~];
(IV) succomber [à ART ~], mourir [de ART ~]
$AntiReal_1$: (IV) réchapper [de ART ~]
$Fact_0$: (I) s'ouvrir, se rouvrir; (II) saigner
$Fact_1$: (IV) emporter, tuer [N]
$Able_1Fact_1$ (≅ Magn) : (I) ouverte < profonde < béante (III) cuisante, douloureuse; (IV) fatale, mortelle, qui ne pardonne pas
$AntiAble_1Fact_1$ (≅ AntiMagn) : bénigne "spéc", sans conséquence

Nous ne pouvons expliquer ici les sens des différentes fonctions lexicales (cf. Mel'čuk *et al.* 1995). Chaque fonction lexicale simple (Magn, $Oper_1$, …) correspond à une règle sémantique particulière (voir Section 3.3.3) et les fonctions lexicales complexes ($IncepOper_1$, $Able_1Fact_1$, …) correspondent à des opérations naturelles sur les fonctions lexicales simples (Kahane & Polguère 2001). Les fonctions lexicales jouent un grand rôle dans les choix lexicaux (Mel'čuk 1988a, Polguère 1998), ainsi que dans la paraphrase et la traduction (Mel'čuk 1988b).

3.3.2 Les modules de correspondance d'un modèle Sens-Texte

La grammaire d'un modèle Sens-Texte est divisée en modules. Chaque module assure la correspondance entre deux niveaux adjacents : le *module sémantique* assure la correspondance entre le niveau sémantique et le niveau syntaxique profond, le *module syntaxique profond* la correspondance entre le niveau syntaxique profond et le niveau syntaxique de surface, le *module syntaxique de surface* la correspondance entre le niveau syntaxique de surface et le niveau morphologique profond, etc.

Les règles de grammaire d'un modèle Sens-Texte sont toutes des *règles de correspondance* entre deux niveaux adjacents, c'est-à-dire des règles qui associent un fragment d'une structure d'un niveau donné avec un fragment d'une structure d'un niveau adjacent. Les règles se présentent toutes sous la forme $A \Leftrightarrow B \mid C$ où A et B sont des fragments de structure de deux niveaux adjacents et C est un ensemble de conditions. La règle doit être lue "si les conditions C sont vérifiées, A peut être traduit par B" dans le sens de la synthèse et "si les conditions C sont vérifiées, B peut être traduit par A" dans le sens de l'analyse. En fait, ce n'est pas l'ensemble des configurations A et B qui sont traduites l'une dans l'autre : les configurations contiennent aussi des éléments qui indiquent comment la règle va s'articuler avec d'autres règles, comment la configuration produite par la règle va s'attacher aux configurations produites par les autres règles (voir Section 3.3.3). Suivant Kahane & Mel'čuk 1999, nous séparons les règles en règles nodales et sagittales : les *règles nodales* sont les règles où la portion de A manipulé par la règle est un nœud, tandis que les *règles sagittales* (lat. *sagitta*) sont les règles où la portion de A manipulée par la règle est une flèche (une dépendance sémantique, syntaxique ou, pour le niveau morphologique, une relation d'ordre).

Nous allons maintenant présenter les trois premiers modules d'un MST.

3.3.3 Le module sémantique d'un modèle Sens-Texte

Le module sémantique réalise la correspondance entre le niveau sémantique et le niveau syntaxique profond. Le module sémantique assure deux opérations fondamentales : la lexicalisation et la hiérarchisation ou arborisation du graphe sémantique.

La hiérarchisation est assurée par les règles sagittales. Parmi les règles sagittales sémantiques, on distingue les règles positives et négatives. Une *règle positive* transforme une dépendance sémantique en une dépendance syntaxique de même direction, tandis qu'une *règle négative* inverse la direction. L'arborisation consiste à choisir une entrée dans le graphe qui donnera la racine de l'arbre, puis à parcourir le graphe à partir de ce nœud d'entrée. Les dépendances sémantiques parcourues positivement (du prédicat vers l'argument) seront traduites par des règles positives, tandis que les dépendances parcourues négativement seront traduites par des règles négatives. Le choix du nœud d'entrée, ainsi que celui des nœuds où seront coupés les cycles du graphe, est guidé par la structure communicative. Nous ne développerons pas ce point ici (cf. Polguère 1990, Kahane & Mel'čuk 1999). Notons simplement que le nœud d'entrée est par défaut le nœud dominant du rhème lorsque celui-ci peut être lexicalisé par un verbe (ou une tournure équivalente de type verbe support-nom prédicatif ou verbe copule-adjectif) et qu'il prend le nœud dominant du thème comme argument sémantique.

Une règle sagittale sémantique positive traduit une dépendance sémantique en une dépendance syntaxique profonde actancielle, tandis qu'une règle négative traduit une dépendance sémantique en une dépendance syntaxique profonde ATTR, COORD ou APPEND (Figure 8). Chaque dépendance sémantique est attachée a deux nœuds sémantiques 'X' et 'Y' dont les correspondants syntaxiques profonds sont X et Y ; ces étiquettes permettent de s'articuler la règle décrite ici avec les règles nodales qui traduisent 'X' en X et 'Y' en Y. La grosse flèche que nous indiquons dans la partie gauche de la règle indique le sens de parcours et doit être compatible avec la structure communicative (cf. Mel'čuk & Kahane 1999).

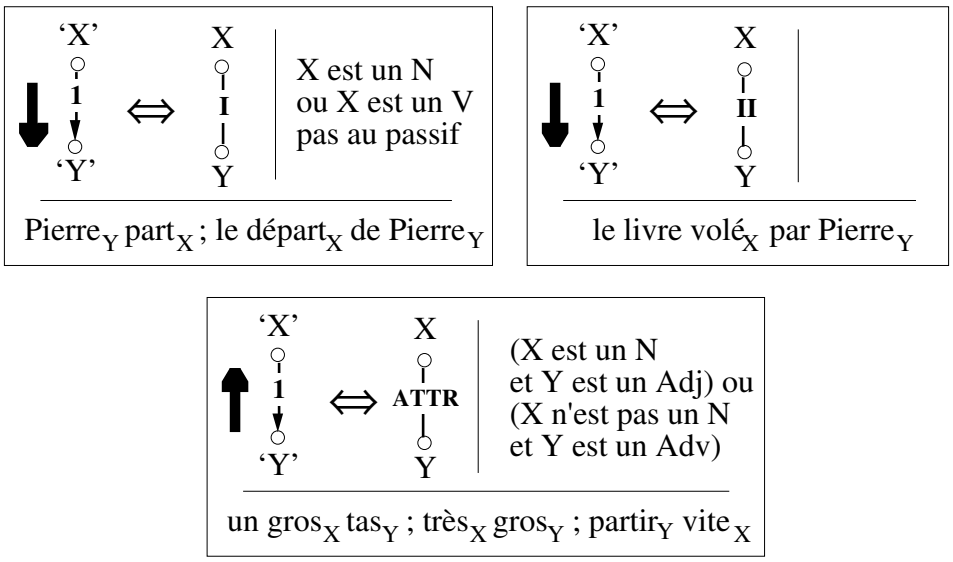

Figure 8 : Trois règles sémantiques sagittales

Toutes les règles que nous présentons dans la Figure 8 sont *locales*, c'est-à-dire que la dépendance syntaxique profonde qui traduit la dépendance sémantique considérée doit être attachée à la traduction nœuds X et Y des nœuds 'X' et 'Y' auxquels est attaché la dépendance sémantique. Il existe pourtant des disparités entre les structures sémantiques et syntaxiques profondes nécessitant des règles non locales (cf. Kahane & Mel'čuk 1999 pour des règles non locales pour le traitement des phrases à extraction).

Les règles sémantiques nodales associent un sémantème à une lexicalisation de ce sémantème. La plupart des sémantèmes sont lexicalisés par une lexie profonde. Certains sémantèmes comme 'intense' vont être lexicalisé par une fonction lexicale (ici Magn), dont la valeur sera recherchée par le module syntaxique profond dans l'entrée lexicale de l'argument concerné. Enfin, des règles sémantiques particulières assurent la réalisation des sémantèmes grammaticaux par des grammèmes de surface.

Terminons notre présentation du module sémantique en montrant comment on passe de la représentation sémantique de (1) (Figure 5) à sa représentation syntaxique profonde (Figure 6). On commence par choisir le nœud d'entrée de la représentation sémantique. Le sémantème 'essayer' est choisi car il est le nœud dominant du rhème, qu'il peut être lexicalisé par un verbe et qu'il prend le nœud dominant du thème comme argument. Ce nœud est lexicalisé par ESSAYER. Ensuite, on parcourt le graphe à partir de ce nœud. Le cycle formé par 'essayer', 'parler' et 'Zoé' sera coupé au niveau de 'Zoé' afin d'assurer la connexité du rhème. Le sens de parcours de toutes les dépendances sémantique est maintenant décidé. Les dépendances sémantiques parcourues positivement donneront des dépendances syntaxiques profondes actancielles. Seule la dépendance entre 'beau' et 'dame', parcourue négativement, donnera une dépendance ATTR. Comme 'dame' sera lexicalisé par le nom DAME, 'beau' devra être lexicalisé par un adjectif. Nous ne présentons pas les règles grammaticales.

3.3.4 Le module syntaxique profond d'un modèle Sens-Texte

Le module syntaxique profond réalise la correspondance entre le niveau syntaxique profond et le niveau syntaxique de surface. Le module syntaxique profond doit assurer l'introduction de toutes les lexies de surface de la phrase (lesquelles correspondent un à un aux mots de la phrase, à l'exception de cas de réduction comme *de le* en *du*).

Les règles syntaxiques profondes sagittales traduisent une dépendance syntaxique profonde en fonction de la nature des éléments qu'elle relient et de leurs tableaux de régime. En particulier, ces règles introduisent les prépositions régies (Figure 9).

Les règles syntaxiques profondes nodales traduisent une lexie profonde. La plupart de ces règles sont contrôlée par le lexique, comme l'expansion d'une locution, ou l'introduction de la valeur d'une fonction lexicale. Les règles syntaxiques profondes nodales comprennent également les règles de pronominalisation : dans une chaîne de référence (c'est-à-dire une chaîne de lexies profonde qui correspondent à un même nœud sémantique), il faut remplacer sous des conditions précises certaines lexies par des pronoms. Ces règles n'ont pas fait l'objet d'une étude sérieuse pour l'instant.

Le module syntaxique profond contient également des règles grammaticales qui assurent la traduction des grammèmes profonds, notamment ceux qui comme la détermination, la voix ou le temps peuvent s'exprimer par des expressions analytiques comprenant des mots. Là encore ces règles n'ont pas fait l'objet d'une étude sérieuse en TST. Nous en proposerons Section 4.2.1 dans le formalisme GUST.

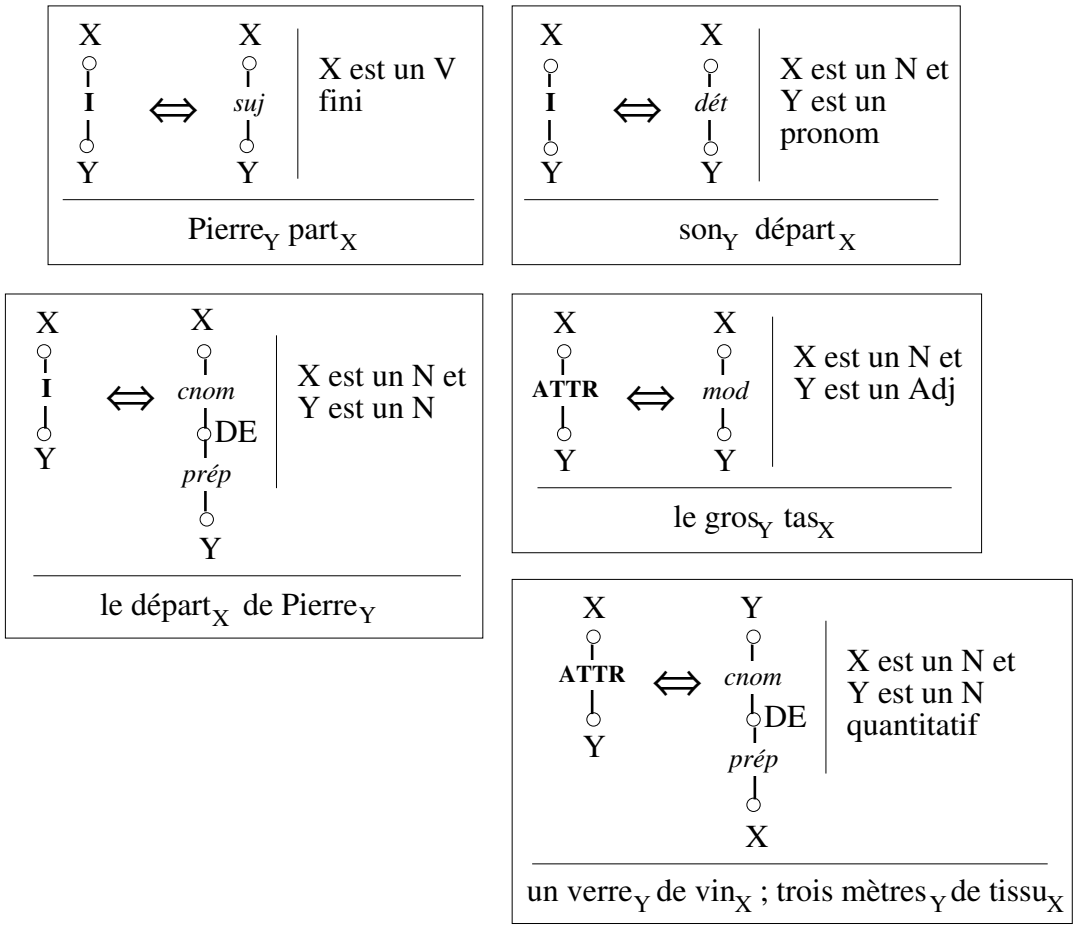

Figure 9 : Cinq règles syntaxiques profondes sagittales

Montrons comment on passe de la représentation syntaxique de (1) (Figure 6) à sa représentation syntaxique de surface (Figure 7). Comme le verbe ESSAYER est fini, l'actant I de ESSAYER devient *sujet*. La règle sagittale qui traduit l'actant II de ESSAYER doit, en fonction du tableau de régime de ESSAYER, introduire une relation syntaxique *infinitive*, la préposition DE et le grammème **infinitif** sur PARLER. L'actant I de PARLER est effacé par la règle de "pronominalisation" de l'actant I d'un verbe à l'infinitif. L'actant II de PARLER est

traduit, en fonction du tableau de régime de PARLER, par la relation d'objet indirect (*iobj*) et la préposition À. Comme BEAU est un adjectif, la relation ATTR donne la relation syntaxique *modifieur*. Enfin, comme DAME n'a pas de déterminant, le grammème **défini** sur DAME donne le déterminant LE, relié à DAME par une relation *déterminative*.

3.3.5 Le module syntaxique de surface d'un modèle Sens-Texte

Le module syntaxique de surface réalise la correspondance entre le niveau syntaxique de surface et le niveau morphologique profond. Le module syntaxique de surface assure la linéarisation, l'accord et le régime (Figure 10) (cf. Mel'čuk & Pertsov 1987 pour un fragment conséquent du module syntaxique de l'anglais).

Les règles de linéarisation indiquent comment un élément se place par rapport à son gouverneur (X < Y ou Y < X). Mais, elles doivent aussi indiquer comment les différents dépendants d'un même nœud se placent les uns par rapport aux autres. Plusieurs techniques sont possibles : on peut par exemple indiquer dans la règle de linéarisation d'un dépendant quels sont les autres dépendants qui peuvent se placer entre lui et son gouverneur (Mel'čuk & Pertsov 1987, Nasr 1996). Nous préférons encoder le placement des co-dépendants par une marque de *position* indiquant la "distance" d'un dépendant donné au gouverneur (Mel'čuk 1967, Courtin & Genthial 1998, Kahane 2000a). Comme métaphore, on peut voir les dépendances comme des élastiques auxquels sont accrochés les mots avec un poids égal à la valeur du trait position : plus le poids est grand (en valeur absolu), plus le mot est loin de son gouverneur. On peut également voir les marques de position comme des adresses de positions précises ouvertes par le gouverneur. Par exemple, un verbe fini ouvre 7 positions devant lui pour les clitiques (*il* < *ne* < *me* < *le* < *lui* < *en* < *y*). Nous donnons Figure 10 les règles de linéarisation du sujet : un sujet non pronominal peut se placer devant le verbe à la position -10 ou après le verbe à la position +10 sous certaines conditions, tandis qu'un sujet pronominal se cliticise et occupe la position -7 devant le verbe. De même, un objet direct pronominal se cliticise et occupe la position -5 ou -4 selon sa personne.

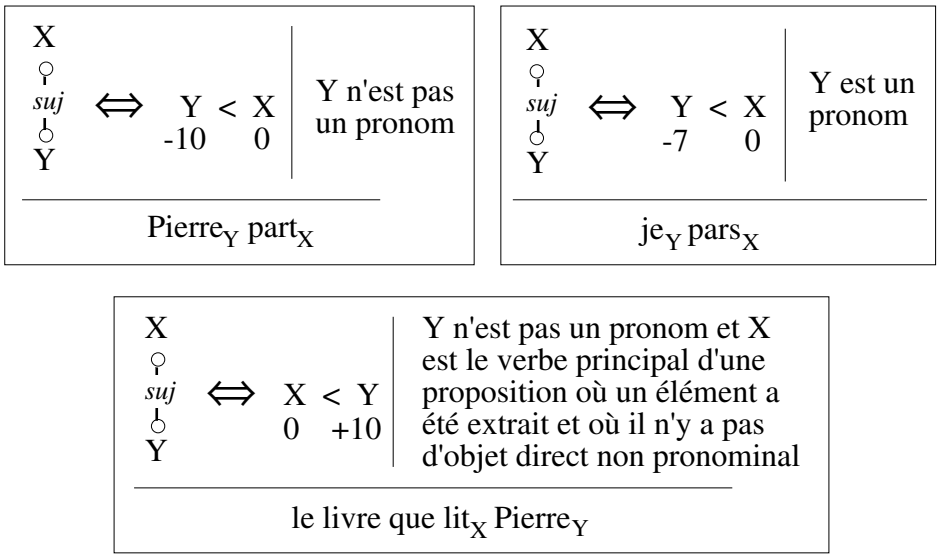

Figure 10 : Trois règles syntaxiques de surface sagittales

Notons que le placement des co-dépendants dépend également de la taille du syntagme dominé par le dépendant (les gros syntagmes ont tendance à être plus éloignés) et de la structure communicative (les syntagmes les plus saillants communicativement ont tendance a être plus éloignés); la marque de position devrait donc être une fonction dépendant de la relation syntaxique, de la taille du syntagme et de la saillance communicative.

Pour le traitement des constructions non projectives, des règles non locales sont nécessaires, puisque l'élément ne se place plus par rapport à son gouverneur, mais par rapport à un ancêtre plus éloigné.[26]

Montrons comment on passe de la représentation syntaxique de surface de (1) (Figure 7) à sa représentation morphologique profonde (2). La racine ESSAYER de l'arbre syntaxique est placée en premier. Le sujet ZOÉ est placé à sa gauche et la préposition DE, qui est la tête de son complément infinitif, à sa droite, conformément aux règles de linéarisation des relations *sujet* et *infinitive*. Le dépendant PARLER de la préposition DE est placé à sa droite, puis la préposition À, qui est la tête de l'*objet indirect* de PARLER, à sa droite, puis le nom DAME qui dépend de À à sa droite. L'article LE et l'adjectif BEAU seront placés à gauche de DAME. En raison de la projectivité, ils devront se placer entre DAME et son gouverneur À. Enfin conformément aux marques de position des règles de placement du *déterminant* et du *modifieur*, l'article LE sera placé à gauche de l'adjectif BEAU.

Nous terminons ici notre présentation de la TST standard. On trouvera une description des règles morphologiques dans Mel'čuk 1993-2001.

4 Une grammaire Sens-Texte basée sur l'unification

Afin de proposer une version complètement formalisée de la TST et d'établir le lien entre l'approche Sens-Texte et d'autres approches, nous allons montrer comment les règles de correspondance d'un modèle Sens-Texte peuvent être interprétées comme des règles génératives basée sur l'unification, c'est-à-dire comment un modèle Sens-Texte standard peut être simulé par une grammaire qui génère des portions de structures et les combine par unification. Le formalisme que nous présentons sera appelé GUST (Grammaire d'Unification Sens-Texte). Nous ferons le lien entre GUST et d'autres formalismes bien connus comme HPSG et TAG, dont il s'inspire d'ailleurs largement.

4.1 Grammaires transductives et grammaires génératives

Les règles de la TST sont des règles qui mettent en correspondance deux fragments de deux structures de deux niveaux de représentation adjacents (par exemple un fragment de structure syntaxique de surface avec un fragment de chaîne morphologique profonde, c'est-à-dire un fragment d'arbre de dépendance avec un fragment d'ordre linéaire). Etant donnés deux ensembles S et S' de structures (graphes, arbres, suites, …), nous appellerons *grammaire transductive* entre S et S' une grammaire G qui met en correspondance des éléments de S et de S' par un ensemble fini de *règles de correspondance* qui mettent en correspondance un fragment d'une structure de S avec un fragment d'une structure de S' (Kahane 2000b). Tous les modules de la TST sont des grammaires transductives. Un modèle Sens-Texte, le modèle d'une langue donnée, est encore une grammaire transductive obtenue par composition des différents modules du modèle.[27]

---

[26] La règle d'inversion du sujet que nous proposons Figure 10 devrait être en fait une règle non locale : le sujet inversé ne se place pas par rapport à son gouverneur, mais rapport au nucléus verbal qui contrôle l'extraction. Cf. Kahane 2000a pour une formalisation.

[27] La composée de deux grammaires transductives ne donnent pas trivialement une grammaire transductive. En effet, si G est une grammaire transductive entre S et S' et G' une grammaire transductive entre S' et S'', on peut construire une grammaire transductive GoG' entre S et S'', mais cette grammaire n'est pas obtenue en composant simplement les régles de G avec les règles de G'. La difficulté vient du fait que les fragments de structure de S' considérés par G ne sont pas forcément les mêmes que ceux considérés par G'. Ainsi le module syntaxique profond de la TST considère comme fragments des portions importantes d'arbre syntaxique de

Remarquons qu'une grammaire transductive G entre *S* et *S′* définit davantage qu'une correspondance entre *S* et *S′*. En effet, pour chaque couple (s,s') de structures appartenant à *S* et *S′* et mises en correspondance par G (c'est-à-dire par des règles de correspondance qui vont associer des fragments de s avec des fragments de s'), G définit aussi des partitions de s et s' (les fragments considérés par les règles) et une fonction $\varphi_{(s,s')}$ entre ces partitions. Nous appellerons cela une *supercorrespondance* (Kahane 2000b). Par exemple, le module syntaxique de surface ne fait pas que mettre en correspondance des arbres de dépendance et des chaînes morphologiques : pour chaque arbre et chaîne en correspondance, il met en correspondance les nœuds de l'arbre avec les éléments de la chaîne (par l'intermédiaire des règles nodales) (Figure 11).

La supercorrespondance entre *S* et *S′* définie par une grammaire transductive est mathématiquement équivalente à l'ensemble des triplets (s,s', $\varphi_{(s,s')}$) où s et s' sont des éléments de *S* et *S′* mis en correspondance et $\varphi_{(s,s')}$ est la fonction associant les partitions de s et s' définies par la mise en correspondance de s et s'. Un triplet de la forme (s,s', $\varphi_{(s,s')}$) est en fait une *structure produit* au sens mathématique du terme, c'est à dire une structure complexe obtenue par l'enchevêtrement de deux structures (l'enchevêtrement est du au fait que, en un sens, les deux structures sont définies sur le même ensemble – l'ensemble des fragments mis en correspondance). Pour prendre l'exemple du module syntaxique de surface d'un MST, si s est un arbre de dépendance, s' une suite et $\varphi_{(s,s')}$ une correspondance entre les nœuds de s et les éléments de s', le triplet (s,s', $\varphi_{(s,s')}$) n'est autre qu'un arbre ordonné (Figure 11).

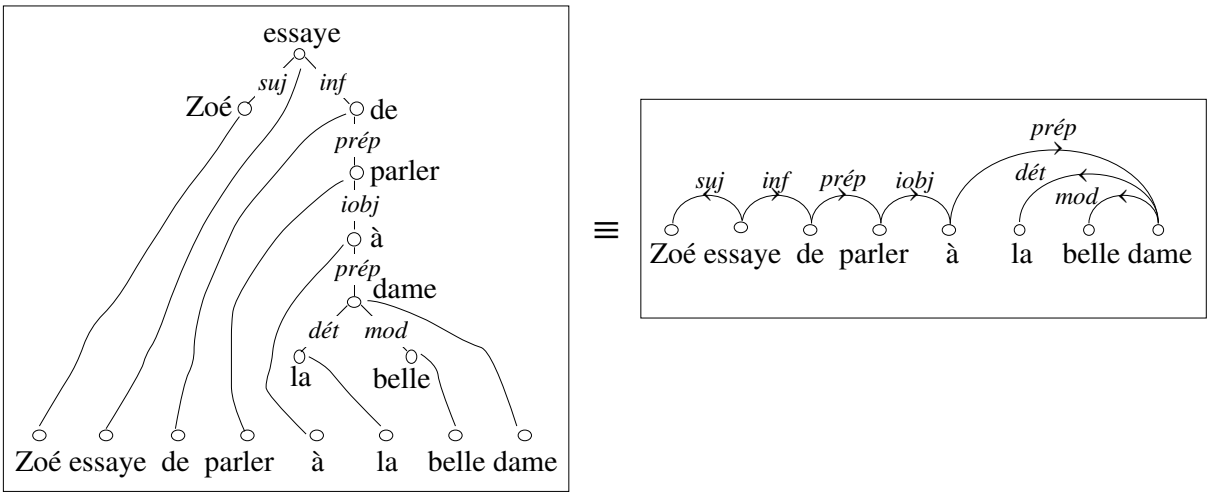

Figure 11 : Equivalence entre un arbre et une suite en correspondance et un arbre ordonné

Une grammaire transductive entre *S* et *S′* peut être simulée par une grammaire générative qui génère l'ensemble des triplets (s,s', $\varphi_{(s,s')}$) décrit par G. Les règles de correspondance sont alors vues comme des règles générant des fragments de structure produit. Nous allons développer cette idée dans la suite.

Inversement, les grammaires génératives qui génèrent des structures produits peuvent vues comme des grammaires transductives. Par exemple, les grammaires de Gaifman-Hays, qui génèrent des arbres de dépendance *ordonnés*, peuvent être vue comme des grammaires mettant en correspondance des arbres de dépendance non ordonnés avec des suites, c'est-à-dire comme une implémentation d'un module syntaxique de surface de la TST. Notons quand même que les

---

surface qui correspondent dans la structure syntaxique profonde à un seul nœud (par exemple pour la règle d'expansion d'une locution), alors que le module syntaxique de surface n'en considère pratiquement pas (et pas les mêmes).

grammaires de Gaifman-Hays cherchent aussi à assurer la bonne formation des arbres de dépendance, alors que, dans le cadre de la TST, celle-ci résulte de l'interaction des différents modules.

4.2 Grammaire d'Unification Sens-Texte

Nous allons maintenant montrer comment un modèle Sens-Texte peut être simulé par une grammaire générative basée sur l'unification, que nous appelons GUST (*Grammaire d'Unification Sens-Texte*).[28] Le formalisme de GUST s'inspire de la grammaire de Nasr (1995, 1996; Kahane 2000a), elle-même inspirée des grammaires TAG lexicalisées (Schabes 1990, Abeillé 1991, XTAG 1995, Candito 1999)..[29]

Nous ne considérons que 3 des 7 niveaux de représentations de la TST : le niveau sémantique, le niveau syntaxique (de surface) et le niveau morphologique (profond). Le 4ème et dernier niveau considéré sera le texte lui-même, c'est-à-dire la séquence des caractères de la phrase. Nous aurons ainsi trois modules (modules sémantique, syntaxique et morphologique). Nous allons les présenter maintenant, puis nous étudierons leurs différentes combinaisons, ce qui nous permettra de faire le lien avec les grammaires complètement lexicalisées comme TAG.

4.2.1 Module sémantique de GUST

Le module sémantique de GUST assure directement la correspondance entre le niveau sémantique et le niveau syntaxique de surface, sans considérer un niveau syntaxique profond intermédiaire. Nous considérons deux types de règles : des *règles sémantiques lexicales*, qui manipulent la configuration sémantique formée d'un sémantème lexical et de ses arguments (plus exactement des dépendances sémantiques vers ses arguments), et des *règles sémantiques grammaticales*, qui manipulent un sémantème grammatical (et la dépendance vers son argument). Ces deux types de règles suffisent à assurer la correspondance entre un graphe sémantique et un arbre syntaxique de surface, puisque n'importe quel graphe sémantique peut être partitionné en un ensemble de configurations prises en entrée par nos règles sémantiques lexicales et grammaticales.

On voit Figure 12 la règle qui donne la réalisation syntaxique de surface de la configuration sémantique composée du prédicat ‘parler’ et de ces deux premiers arguments. Dans la TST standard, cette information se trouve dans l'entrée de dictionnaire de PARLER. En un sens, le dictionnaire ne dit pas comment obtenir la correspondance entre ces deux configurations, et l'information de la Figure 12 est en fait le résultat de la composition de plusieurs règles nodales et sagittales sémantiques et syntaxiques profondes déclenchées sous le contrôle de l'entrée de dictionnaire de PARLER. Dans la règle de la Figure 12, il est également indiqué que PARLER est un verbe et que ce verbe doit recevoir des grammèmes profonds de mode, temps et voix. Les flèches ($\rightarrow$) qui précèdent ces grammèmes indiquent que ceux-ci ne sont pas encore exprimés. Ils le seront par des règles sémantiques grammaticales qui seront obligatoirement déclenchées (en exigeant que les flèches aient disparues dans une représentation syntaxique bien formée).

---

[28] GUST n'est pas exactement une autre présentation de la TST. Certains choix théoriques peuvent être différents et nous pensons que ce formalisme permet de résoudre certaines questions dont le traitement classique en TST n'est pas très clair, notamment tout ce qui concerne l'interaction entre les différentes règles d'un même module ou de deux modules adjacents.

[29] Voir également Hellwig 1986 pour une proposition antérieure de grammaire de dépendance basée sur l'unification, appelée DUG (Dependency Unification Grammar).

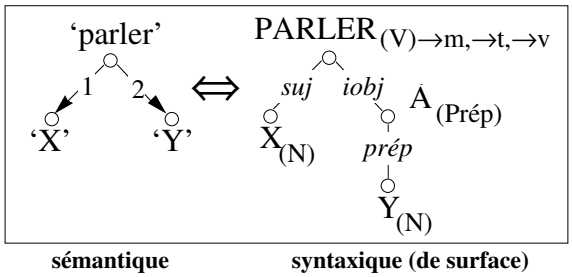

Figure 12 : Une règle de correspondance TST (sémantico-syntaxique profonde)

Nous proposons Figure 13 la règle de GUST qui simule la règle TST de la Figure 12. Au lieu de mettre en correspondance un fragment de structure sémantique avec un fragment de structure syntaxique, cette règle propose un fragment de structure produit sémantique-syntaxique, exprimant à la fois la relation entre le sémantème 'parler' et la lexie PARLER et les relations entre les arguments 1 et 2 de 'parler' et les *sujet* et *objet indirect* (*iobj*) de PARLER. Dans la règle de la Figure 13, l'arbre syntaxique est représenté explicitement (ce qui donne une certaine primauté à la syntaxe), alors que le graphe sémantique est encodé dans l'étiquetage des nœuds par l'intermédiaire des traits sém, arg1 et arg2. Chaque nœud possède un trait sém dont la valeur est un sémantème, le signifié de la lexie étiquetant ce nœud. Lorsque ce sémantème a des arguments, ceux-ci sont les valeurs des traits arg1, arg2, etc., valeurs qui sont partagées avec les traits sém des nœuds syntaxiques qui réalisent ces arguments. Le partage d'une valeur par plusieurs traits est indiqué par une variable, la valeur elle-même n'étant indiquée qu'une fois.[30] Notons encore que les mots vides portent un trait ¬sém qui bloquera l'unification avec une étiquette portant le trait sém.

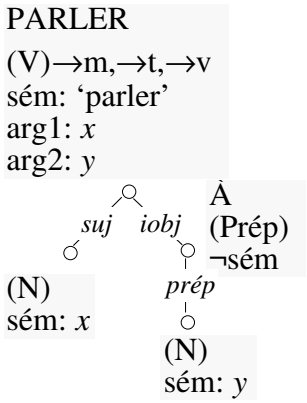

Figure 13 : Une règle sémantique lexicale GUST

Les règles lexicales se combinent par unification. Nous présentons Figure 14 la dérivation de la phrase *Le petit chat dort ici* par combinaison des règles lexicales associées aux lexies de cette phrase (il s'agit en fait de règles lexicales sur lesquelles ont déjà été appliquées les règles grammaticales, comme on le verra plus loin). Deux règles se combinent par fusion de deux nœuds et unification des étiquettes correspondantes. Comme nous le verrons dans la suite, plusieurs nœuds, ainsi que des dépendances, peuvent fusionner lors de la combinaison de deux règles. Le résultat d'une dérivation est bien formé si cette dérivation met bien en correspondance

---

[30] Le fait de faire partager une même valeur à plusieurs traits est une technique bien connue dans les formalismes basés sur l'unification. Voir l'usage intensif qu'en fait par exemple le formalisme HPSG (Pollard & Sag 1994).

un graphe sémantique *connexe* avec un arbre de dépendance, c'est-à-dire si le résultat est un arbre dont tous les traits sém sont instanciés (certains traits ont pour valeur une variable qui indique en fait l'adresse de la valeur d'un autre trait qui est instancié).

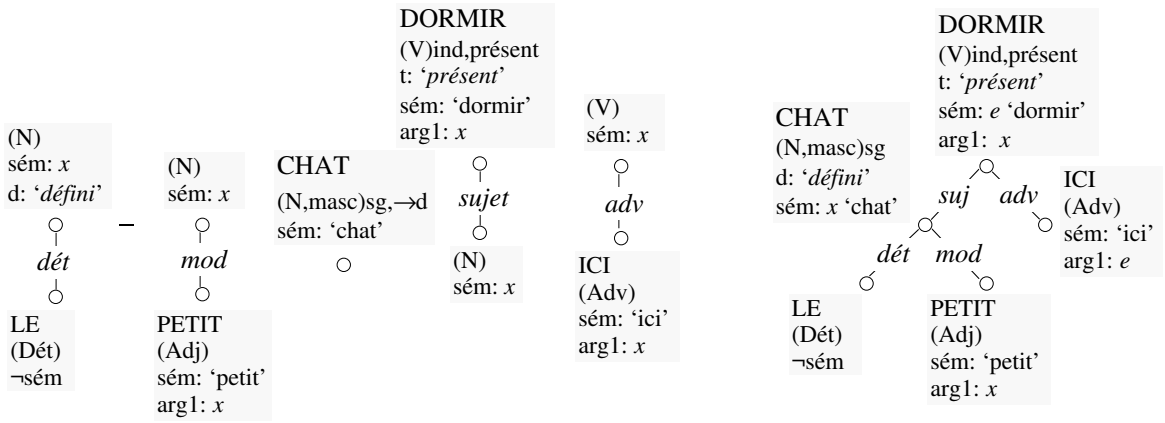

Figure 14 : Dérivation de *Le petit chat dort ici*

Avant de revenir sur les règles lexicales, nous allons présenter les règles sémantiques grammaticales. Les grammèmes profonds sont calculés à partir de la représentation sémantique : certains yapparaissent explicitement comme des sémantèmes grammaticaux, d'autres seront calculés à partir de la structure communicative sémantique (comme la voix qui dépend en partie de la partition thème-rhème) et d'autres encore sont imposés par la rection (comme le mode infinitif). Il n'est pas aisé de traiter la combinaison entre une règle lexicale et une règle grammaticale par unification, car un grammème ne fait pas qu'ajouter de l'information : il peut aussi entraîner une modification importante du comportement de la lexie qu'il spécifie. C'est le cas, par exemple, d'un grammème de voix passive qui entraîne une redistribution des fonctions des actants syntaxiques de la lexie (l'*objet* devient *sujet* et le *sujet* devient un complément d'agent).

Dans un premier temps, nous allons traiter les règles grammaticales comme des opérateurs qui associent à une règle lexicale une nouvelle règle lexicale (où un grammème profond supplémentaire est exprimé). Nous présentons Figure 15 les règles pour le **présent**, le **passé composé** et la voix **passive**. Le *passé composé* d'un verbe *X* est exprimé par l'auxiliaire AVOIR[31] au **présent** et le verbe *X* au **participe passé**. L'auxiliaire AVOIR est l'auxiliaire par défaut : si *X* possède un trait aux indiquant un autre auxiliaire (par exemple ÊTRE), la valeur @aux de ce trait sera utilisée à la place de AVOIR. [32] Le sémantème 'passé composé' apparaît dans l'étiquette de *X* (son argument est la valeur du trait sém de *X*), mais le grammème profond **passé composé** n'apparaît pas en tant que tel. Seuls apparaissent les grammèmes de surface tels que **présent** ou **p-passé** (participe passé). Notez également le positionnement des traits pour le mode sur l'auxiliaire et de la voix sur le verbe X.

---

[31] Nous passons sous silence la question de la sémantique de l'auxiliaire. Nous devons assurer que les modifieurs de la forme verbale composée qui dépendent sémantiquement de l'auxiliaire prennent bien le signifié du verbe comme argument sémantique.

[32] La notation a//b utilisée dans les règles comme valeur d'un trait signifie : la valeur est a, si la valeur b ne peut être trouvée, et b sinon.

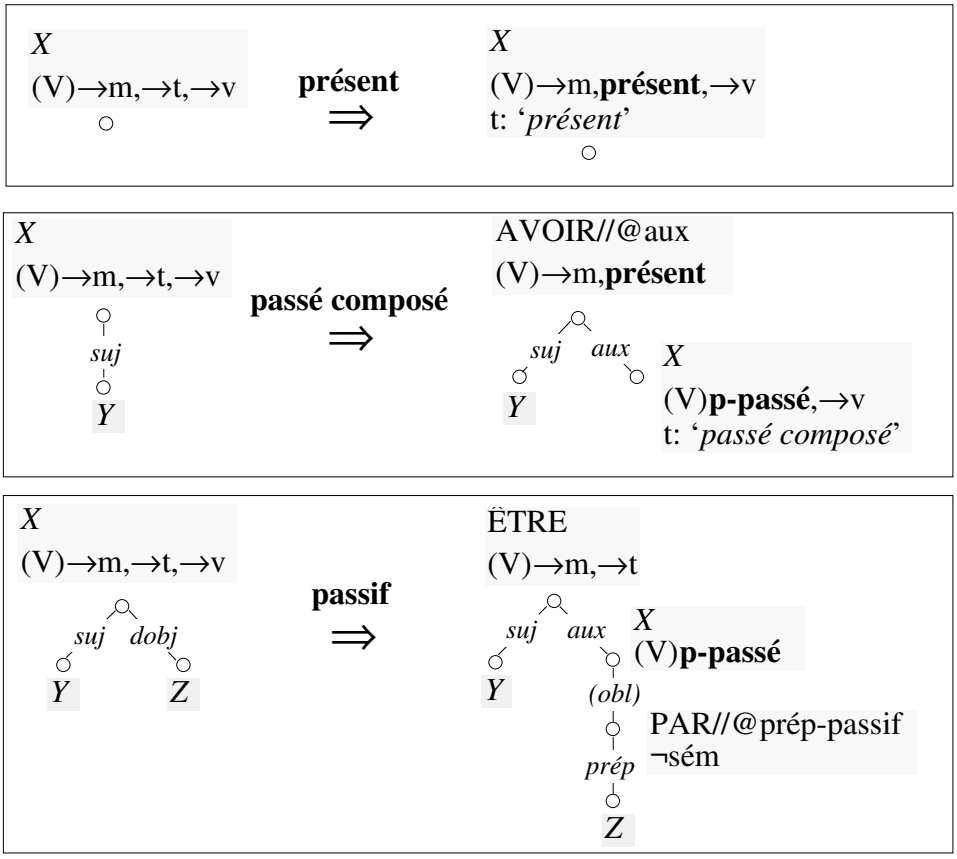

Figure 15 : Règles grammaticales GUST pour le **présent**, le **passé composé** et la voix **passive** (version opérateur)

Nous proposons Figure 16 une autre version de la règle grammaticale pour le **passé composé**. Cette règle se combine par unification avec la règle lexicale d'un verbe. La montée du sujet *Y* de *X* sur l'auxiliaire est assurée par la flèche étiquetée *sujet* de *X* à *Y*. Cette flèche, que nous appelons une *quasi-dépendance*, va fusionner avec la dépendance *sujet* de *X* (dans sa règle lexicale) et tuer cette dépendance. Nous proposons également une règle grammaticale pour le **défini** lorsqu'il est exprimé par l'article LE. La détermination (**défini**, **indéfini**, **partitif**) est un grammème profond qui a une expression purement analytique et ne donne donc pas de grammème de surface.[33]

[33] Ce comportement marginal de la détermination peut pousser certains à ne pas traiter la détermination comme une catégorie flexionnelle et à préférer traiter les lexies LE ou UN comme des lexies pleines exprimant les sens 'défini' et 'indéfini'. Nous préférons notre solution. Cette solution, par l'obligation d'exprimer la détermination, règle aussi le problème de la présence obligatoire du déterminant.

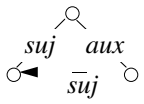

Figure 16 : Règles grammaticales GUST pour le **passé composé** et le **défini** (version unification)

Nous allons maintenant montrer comment sont traités quelques phénomènes linguistiques en proposant d'autres règles sémantiques lexicales. Nous donnons Figure 17 la règle pour la locution LA MOUTARDE MONTER AU NEZ. La règle sémantique pour une locution fait correspondre un sémantème à une configuration de lexies de surface. Dû au fait que seul la racine de cette configuration accepte des modifications (cf. (4)), seule la racine de l'arbre aura un trait sém (instancié par le signifié de locution), tandis que les autres nœuds auront un trait ¬sém qui bloquera toute modification (puisqu'un modifieur est un prédicat qui prend son gouverneur comme argument et exige donc que celui-ci ait un trait sém (cf. les règles pour PETIT et ICI de la Figure 14). Un verbe avalent (sans argument) comme PLEUVOIR aura une règle similaire à celle d'une locution, avec un nœud ¬sém pour le sujet vide.[34]

(4) a. *La moutarde me monte **sérieusement** au nez*
b. **La moutarde **forte** me monte au nez*

Figure 17 : Règles lexicales pour une locution et pour un verbe avalent

Le contraste entre verbe à contrôle (comme ESSAYER) et verbes à montée (comme COMMENCER) est traditionnellement encodé dans les grammaires syntagmatiques dans la structure syntaxique. Dans notre approche, les deux types de verbes ont exactement la même représentation syntaxique : le verbe gouverne un sujet et un infinitif qui partage avec le verbe le même sujet (nous reviendrons sur la relation *sujet* de l'infinitif). Le contraste vient de la

---

[34] Le traitement est différent en TST où l'introduction d'un sujet vide résulte d'une règle grammaticale syntaxique profonde. D'ailleurs, notre traitement n'est pas entièrement satisfaisant. Il serait probablement préférable de traiter le sujet de PLEUVOIR comme un élément grammatical et non comme une portion de locution, puisque celui-ci n'apparaît pas dans certaines constructions comme *Dieu fait pleuvoir*.

représentation sémantique : un verbe à contrôle prend son sujet comme argument sémantique, mais pas un verbe à montée (Figure 18). Dans les deux cas, l'infinitif contrôle le sujet de son gouverneur et il faut un moyen d'assurer cela. Pour cela, nous considérons qu'un infinitif possède une sorte de dépendance *sujet* ; ce lien s'apparente à une dépendance, mais n'en est pas une, car il ne compte pas dans la structure d'arbre et il n'est pas pris en compte dans la linéarisation (cf. Hudson 2000 pour une proposition similaire). Un tel lien sera appelé une *quasi-dépendance*. De même, on aura dans la règle d'un verbe à contrôle ou à montée une quasi-dépendance *sujet* pour l'infinitif avec laquelle la quasi dépendance de la règle de l'infinitif devra s'unifier. La quasi-dépendance est donc juste un moyen assez simple d'assurer le contrôle du sujet du verbe à contrôle ou à montée par l'infinitif. Notons que ce contrôle est bien syntaxique : il s'agit du sujet du verbe infinitif et pas d'un argument sémantique précis. En effet, il peut s'agir d'un sujet sémantique vide (5a), d'un sujet qui fait partie d'une locution (5b) et, lorsque ce sujet est plein, il peut s'agir aussi bien du premier argument (5c) que du second argument (5d). En conséquence, les infinitifs doivent avoir un sujet dans leur représentation syntaxique, mais ce sujet sera une quasi-dépendance (afin d'éviter qu'un verbe infinitif ait un vrai sujet). La règle sémantique du grammème **infinitif** devra assurer que la relation *sujet* devienne bien une quasi-dépendance .

(5) a. *Il commence à pleuvoir.*
b. *La moutarde commence à lui monter au nez.*
c. *Le bruit commence à géner le garçon.*
d. *Le garçon commence à être gené par le bruit.*

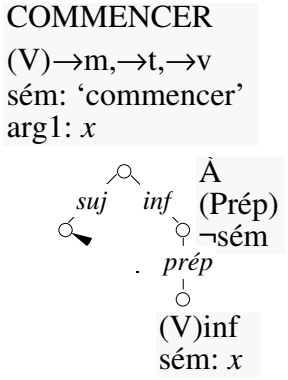

Figure 18 : Règles lexicales pour un verbe à contrôle et un verbe à montée

Remarquons que le fait qu'un verbe à contrôle prenne son sujet comme argument sémantique suffit à éviter que ce verbe ait un sujet vide (ce qui est un contraste bien connu entre verbes à contrôle et verbes à montée) :

(6) a. **Il essaye de pleuvoir.*
b. **La moutarde essaye de lui monter au nez.*

Les verbes copules, c'est-à-dire les verbes prenant un attribut, seront traités de façon similaire aux verbes à montée : dans la règle sémantique lexicale d'un verbe copule, on aura une quasi-dépendance *modifieur* indiquant le lien entre le dépendant du verbe copule “modifié” par l'adjectif attribut et l'adjectif lui-même (Figure 19). Cette quasi-dépendance permet à la fois à l'adjectif de récupérer son argument sémantique et d'assurer l'accord de l'adjectif avec le nom “modifié” par la règle d'accord ordinaire de l'adjectif avec le nom qu'il modifie (et sans qu'il soit nécessaire de faire circuler de l'information au travers du verbe copule). Enfin, cette solution permet d'utiliser la même règle lexicale pour l'adjectif qu'il soit épithète (7a) ou qu'il contrôle le sujet (7b) ou l'objet (7c) (voir Figure 19 la règle pour l'adjectif PETIT).

(7) a. *un petit livre*
b. *ce livre est petit*
c. *Pierre trouve ce livre petit*

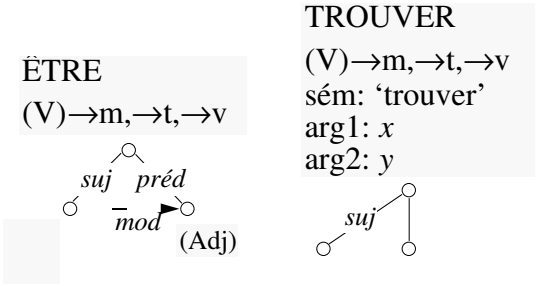

Figure 19 : Règles lexicales pour les verbes copules et les adjectifs

Le phénomène dit du *tough*-movement peut être décrit de la même façon. Quand un adjectif tel que FACILE gouverne un verbe, le nom que modifie l'adjectif n'est pas son argument sémantique mais un argument sémantique du verbe (8a). De plus, le nom doit remplir le rôle d'objet direct du verbe. Par conséquent, la règle sémantique de FACILE contient une quasi-dépendance *objet direct* entre le verbe gouverné et le nom modifié (Figure 19). Ainsi seul un verbe pourvu d'un objet direct peut se combiner avec FACILE et l'"extraction" de l'objet direct du verbe sera assurée par l'unification de la dépendance *objet direct* du verbe avec la quasi-dépendance de même rôle de la règle de FACILE. On peut remarquer que la règle de FACILE peut aussi se combiner avec un verbe copule (8b,c).

(8) a. *un livre facile à lire*
b. *ce livre est facile à lire*
c. *Pierre trouve ce livre facile à lire*

Nous arrêterons là la présentation du module sémantique de GUST. Comme on l'a vu, l'un des objectifs de GUST est d'éviter la multiplication des règles associé à une lexie. Sur le fragment de grammaire proposé, on a pu couvrir avec une seule règle lexicale par lexie un grand nombre de constructions diverses. De même, par la combinaison avec les règles grammaticales, on construit les règles des différentes formes d'un verbe à partir d'une seule règle lexicale.[35]

4.2.2 Module syntaxique de GUST

Le module syntaxique de GUST correspond au module syntaxique de surface de la TST : il assure la correspondance entre le niveau syntaxique de surface et le niveau morphologique profond. Le module syntaxique possède trois types de règles : des règles d'accord, des règles de régime et des règles de linéarisation. On voit Figure 20 la règle d'accord du verbe avec son sujet et la règle de rection des pronoms *sujet* (qui reçoivent le **nominatif**). La règle d'accord du sujet indique que le sujet s'accorde en nombre et en personne avec son sujet (9a). Lorsque le

---

[35] Nous n'avons pas discuté des différentes sous-catégorisations d'une même lexie (par exemple, *demander* N *à* N, *que* V_{subj}, *à* V_{inf}, …). Tel que nous avons présenté le formalisme, nous devrions introduire une règle pour chaque sous-catégorisation. Pour des questions d'efficacité de l'analyse automatique ou de pertinence cognitive (cf. Section 5.3), nous pensons préférable, tant que cela est possible, de rassembler ces différentes sous-catégorisations dans une même règle. Nous devrons alors de introduire des disjonctions et considérer des dépendances optionnelles.

sujet ne possède pas de trait de personne, la valeur par défaut est **3** (9b,c) et lorsqu'il ne porte pas de grammème de nombre la valeur par défaut est le singulier (**sg**) (9c).

(9) a. ***Nous*** *viendrons.*
b. ***Pierre*** *viendra.*
c. ***Que*** *tu viennesest impossible.*

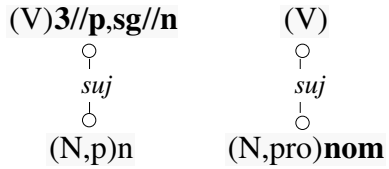

Figure 20 : Règles syntaxiques d'accord et de rection

Les règles de linéarisation de GUST simulent les règles de linéarisation de la TST (présentées dans la Figure 7 de la Section 3.2.3). Nous reprenons Figure 21 la règle de placement par défaut d'un sujet non pronominal. Une règle de ce type met en correspondance la dépendance entre deux nœuds syntaxiques avec une relation d'ordre (agrémentée d'un trait de position) entre les nœuds morphologiques correspondants. Pour préparer le passage à GUST, nous avons déplacé les conditions d'application de la règle.

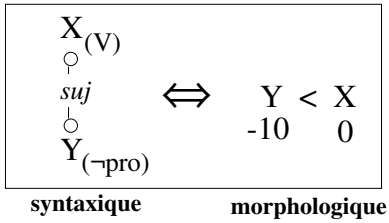

Figure 21 : Une règle de linéarisation TST

Nous proposons Figure 22 la règle de GUST qui simule la règle TST de la Figure 21. Au lieu de mettre en correspondance un fragment de structure syntaxique avec un fragment de structure morphologique, cette règle propose un fragment de structure produit syntaxique-morphologique, c'est-à-dire un morceau d'arbre ordonné.

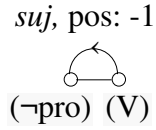

Figure 22 : Une règle de linéarisation GUST

Les règles comme celles de la Figure 22 se combinent par unification. On impose que le résultat soit un arbre ordonné projectif (Kahane 2000b, 2001). Nous ne traiterons pas ici la question de la linéarisation des arbres non projectifs (voir, par exemple, Bröker 1998, Lombardo & Lesmo 1998, Kahane *et al.* 1998, Hudson 2000, Gerdes & Kahane 2001).

4.2.3 Module morphologique GUST

Nous terminons notre présentation des modules de GUST par le module morphologique qui assure la correspondance entre le niveau morphologique (profond) et le niveau textuel, c'est-à-dire la chaîne de caractères qui forme le texte d'une phrase. Les règles TST permettent de mettre

en correspondance la représentation morphologique profonde d'un mot, une lexie de surface accompagnée d'une liste de grammèmes de surface, avec une chaîne de caractères. Nous présentons Figure 23 une règle de ce type dans le style TST [36].

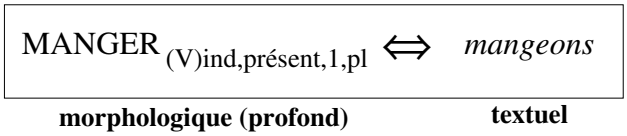

Figure 23 : Une règle morphologique TST

Cette règle est simulée en GUST par une règle qui présente ces deux informations dans une même structure (Figure 24).

Figure 24 : Une règle morphologique GUST

Contrairement à la TST, GUST n'utilise pas de dictionnaire séparé : par exemple, le tableau de régime est complètement encodé dans les règles sémantiques. De même, la partie du discours et tous les traits pertinents (genre des noms, personne des pronoms, comportements particuliers, …) devront être introduit par la règle morphologique.

4.3 Combinaison des modules

Nous allons maintenant montrer comment les règles des différents modules se combinent pour dériver une phrase, c'est-à-dire pour mettre en correspondance une représentation sémantique avec un texte. L'avantage de GUST, sur un modèle Sens-Texte standard, est que, comme pour tous les formalismes basés sur l'unification, il est très facile de combiner n'importe quelles règles ensemble. En particulier, comme nous allons le voir, une grammaire GUST peut garder une forme modulaire, comme la TST, ou être complètement lexicalisée, comme TAG (avec des avantages sur cette dernière, notamment le fait qu'on peut éviter l'explosion du nombre de structures élémentaires associées à chaque entrée lexicale.

4.3.1 Dérivation d'une phrase

Nous présentons Figure 25 l'ensemble des règles nécessaires à la dérivation de la phrase (10) :

(10) *Nous essayons de manger la soupe.*

Ces règles permettent de mettre en correspondance la représentation sémantique de (10) avec le texte de (10). Il y a plusieurs façons d'utiliser ces règles, dans le sens de l'analyse comme de la synthèse. Nous allons regarder le sens de l'analyse. Il s'agit de construire une représentation sémantique à partir du texte. On peut distinguer deux stratégies principales : la stratégie horizontale et la stratégie verticale. La métaphore horizontal/vertical s'entend par rapport au

---

[36] Dans la TST standard, une telle règle est en fait le résultat de la composition d'un grand nombre de règles morphologiques et phonologiques. Si nous voulons être capable de traiter des mots inconnus, nous devrons avoir des règles de ce type.

découpage de l'ensemble des règles de la Figure 25, selon qu'il est fait en tranches horizontales ou verticales.

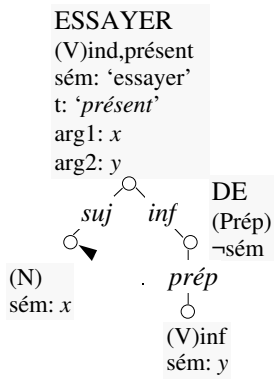

Figure 25 : Dérivation de *Nous essayons de manger la soupe*

4.3.2 Stratégie horizontale

La *stratégie horizontale* consiste à déclencher les règles module après module.

1) Le *module morphologique* permet de passer du texte proprement dit (la chaîne de caractère) à la représentation morphologique, c'est-à-dire une suite de lexies accompagnées d'une liste de grammèmes. La règle introduit également la partie du discours et tous les traits pertinents pour la suite. Le module morphologique réalise ce qu'on appelle traditionnellement la *lemmatisation*, l'*étiquetage morphologique* ou le *tagging*. A noter que le module n'a pas le pouvoir, comme le font ce qu'on nomme généralement des taggeurs, de filtrer certaines séquences de lexies (ou de catégories lexicales) qui ne peuvent apparaître dans la langue. De tels filtres sont en fait la projection d'informations contenues dans le module syntaxique, et nous pensons qu'il est préférable d'utiliser le module syntaxique lui-même pour cette tâche. Notre étiqueteur ne fait donc que proposer pour chaque mot toute les lemmatisations possibles sans tenir compte des étiquettes attribuées aux lexies voisines.

2) Le *module syntaxique* permet de passer de la représentation morphologique à la représentation syntaxique. Il propose pour chaque couple de lexies, en fonction de leurs positions relatives, une liste (éventuellement vide) de dépendances susceptibles de les lier. Nous verrons Section 5 différentes procédures pour produire des arbres syntaxiques. Le module syntaxique réalise ce qu'on appelle traditionnellement le *shallow parsing* ou *analyse*

superficielle. A noter que le module syntaxique n'a pas le pouvoir de contrôler la sous-catégorisation des lexies, ni même de vérifier qu'un verbe a bien un et un seul sujet. Ceci sera contrôlé par le module sémantique. Comme précédemment, il est possible de projeter une partie du module sémantique sur le module syntaxique pour assurer ces points, bien que nous pensions qu'il est préférable d'utiliser le module sémantique lui-même.[37]

3) Le *module sémantique* permet de passer de la représentation syntaxique à la représentation sémantique. Chaque dépendance syntaxique doit être associée à une configuration mise en correspondance avec une configuration sémantique de relations prédicat-argument entre sémantèmes pour être validée. Comme on l'a dit, notre représentation sémantique est une représentation du sens purement linguistique et n'a pas l'ambition d'être une représentation de l'état du monde dénotée par la phrase. Pour cette raison, une grande partie de ce que réalise notre module sémantique est considérée par beaucoup comme une étape de l'analyse syntaxique et correspond à ce qu'on appelle généralement l'*analyse profonde*, ou *deep analysis*.

La stratégie horizontale est la stratégie retenue par la plupart des approches modulaires. Le principal inconvénient de la stratégie horizontale est le fait que la désambiguïsation (quand elle est possible) n'intervient qu'au niveau sémantique et qu'il faudra manipuler aux niveaux morphologique et syntaxique un très grand nombre d'analyses concurrentes.

4.3.3 Stratégie verticale et lexicalisation complète

La stratégie verticale consiste à déclencher les règles mot après mot.

Prenons l'exemple (10). Lorsqu'on analyse le mot *nous*, le module morphologique propose (parmi d'autres propositions) d'étiqueter *nous* comme une forme nominative du pronom NOUS. Mais, on peut alors, par la règle syntaxique de rection, en déduire qu'il s'agit d'un clitique sujet, puis par la règle de linéarisation des pronoms *sujet* et par la règle d'accord prédire la position du verbe et en partie sa forme. La règle sémantique associée à la lexie NOUS peut être également déclenchée. Par la seule analyse du mot *nous*, on peut donc déclencher 5 règles et débuter l'analyse syntaxique et sémantique. La même chose lorsqu'on analyse le mot suivant *essayons*. Le module morphologique proposer d'étiqueter *essayons* comme une forme du verbe ESSAYER. Cette forme remplit les conditions imposées par *nous* à son gouverneur syntaxique et la règle peut donc être immédiatement combinée avec les règles précédentes. Une des règles sémantiques associées à ESSAYER peut être déclenchée. Si l'on déclenche la règle où le deuxième argument est réalisé par DE Vinf, on déclenchera les règles syntaxiques de linéarisation associées à une telle construction. Et ainsi de suite. La Figure 26 montre les différents paquets de règles déclenchés par les différents mots de la phrase.

---

[37] Pour que l'analyseur soit robuste, la grammaire doit proposer un traitement par défaut des mots inconnus. Par exemple, le module sémantique doit proposer, parmi ses différentes règles par défaut, une règle pour une forme verbale qui indiquera que la forme en question doit avoir un sujet et qu'elle aura au plus un objet direct, un objet indirect et deux compléments obliques. Ce sont les projections de ces règles sémantiques par défaut qui sont les règles filtres généralement utilisées par les modules syntaxiques.

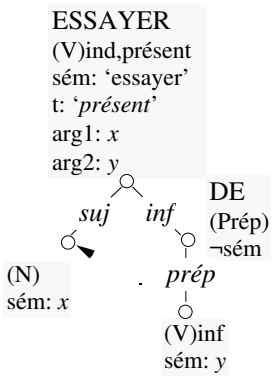

Figure 26 : Regroupement des règles dans l'analyse verticale

Une stratégie d'analyse verticale semble beaucoup plus séduisante qu'une stratégie d'analyse horizontale si l'on se place du point de vue cognitif, c'est-à-dire du point de vue de la modélisation du processus d'analyse linguistique par un locuteur. Même du point de vue du traitement informatique, une analyse verticale pourrait s'avérer plus efficace qu'une analyse horizontale. Il existe d'ailleurs une variante de l'analyse verticale qui consiste à précompiler les paquets de règles déclenchés par chaque mot. On obtient ainsi une grammaire dite *complètement lexicalisée* (*fully lexicalized grammar*). La grammaire ainsi obtenue s'apparente à la grammaire proposée par Nasr 1995, 1996, elle-même inspirée des TAG (cf. également Kahane 2000a pour un traitement des extractions). Le passage à des règles complètement lexicalisées se fait simplement par combinaison d'un paquet de règles modulaires (Figure 27). Il s'agit de la combinaison ordinaire des règles de la grammaire (basée sur l'unification) ; la seule différence est que la combinaison des règles ne se fait pas au moment de l'analyse, mais dans une phase préalable de *précompilation* (Candito 1996, Candito & Kahane 1998).

Le passage d'une grammaire modulaire à une grammaire complètement lexicalisée amène plusieurs commentaires.

1) L'analyse verticale avec la grammaire modulaire revient en fait à utiliser une grammaire complètement lexicalisée sans l'avoir lexicalisée au préalable, mais en construisant les règles lexicalisées à la demande (on line) au moment de l'analyse. Quels sont alors les avantages ou les inconvénients de la grammaire complètement lexicalisée ? La précompilation consomme de l'espace, puisqu'il faut mémoriser toutes les règles lexicalisées, lesquelles sont extrêmement redondantes entre elles. Par contre, le temps d'analyse, si l'accès aux règles compilées est bien géré, devrait être amélioré par le fait qu'une partie des combinaisons de règles est déjà faite. L'alternative entre grammaire modulaire et grammaire complètement lexicalisée peut aussi être considérée du point de vue cognitif : sous-quelle forme la grammaire est-elle codée dans notre cerveau ? Y-a-t-il des

constructions linguistiques plus fréquentes que d'autres qui sont déjà "lexicalisées" ? La grammaire s'acquiert-elle sous forme modulaire ou "lexicalisée" ?

2) Les deux analyses, avec ou sans précompilation, posent les mêmes problèmes théoriques, à savoir quels sont les paquets de règles qui doivent être associés à une lexie donnée ou, de manière équivalente, mais en se plaçant du point de vue des règles plutôt que des lexies, à quelle lexie doit être associée une règle donnée. Prenons un exemple : à quelle lexie, gouverneur ou dépendant, doit être associée une règle de linéarisation? Considérons le cas de l'objet direct en français. Les règles sont les suivantes : un nom *objet direct* se place derrière le verbe, un pronom clitique se place devant le verbe (à une place bien précise par rapport aux autres clitiques) et un pronoms relatif ou interrogatif se place à l'avant de la proposition. Il serait peu économique d'indiquer pour chaque nom, dans la règle lexicalisée qui lui correspond, comment il se place quand il est *objet direct*, *sujet* ou autre chose encore. Il est donc préférable d'attacher la règle de linéarisation de l'objet direct aux verbes qui en possède un. Par contre, le pronom clitique *objet direct* a une forme bien particulière et un placement bien particulier. Il semble plus économique d'attacher à ce seul mot, *le*, le pronom clitique objet direct, les règles qui lui sont spécifiques. De même, les pronoms relatifs ou interrogatifs ont un placement particulier qui ne dépend pas réellement de leur fonction. Il semble donc aussi plus économique que la règle de placement de ces éléments leur soit attachée. La solution retenue est donc de panacher l'information sur le placement de l'objet direct entre les verbes transitifs pour les éléments canoniques (les noms) et les éléments non canoniques eux-mêmes pour ce qui les concerne (voir Figure 27). Cette façon de faire permet d'éviter la multiplication des règles lexicalisées associées à un même verbe, comme cela est le cas par exemple en TAG où le formalisme ne permet pas d'encoder le placement des arguments d'une lexie ailleurs que dans la règle (appelée structure élémentaire en TAG) de cette lexie. Reste une difficulté : il faut éviter, lors de la combinaison d'un verbe x avec un élément en position non canonique y, que rentre en conflit la règle de linéarisation des éléments en position canonique attachée à x avec la règle de linéarisation spécifique attachée à l'élément y (voir Kahane 2000a pour une solution basée sur l'unification consistant à "tuer" la règle de positionnement attachée à la dépendance objet du verbe en l'unifiant avec un leurre, une quasi-dépendance objet placée dans la structure associée à l'élément y).

3) En dehors des questions computationnelles, les grammaires complètement lexicalisées ont un autre intérêt : il est très facile d'écrire un premier fragment de grammaire et de l'étendre à chaque nouvelle construction rencontrée. Néanmoins, de cette façon, on contrôle difficilement la consistance globale de la grammaire et certaines constructions obtenues seulement par combinaison de phénomènes divers peuvent être facilement oubliées (par ex., la forme passive d'un verbe où un complément est relativisé et un autre cliticisé). Pour cette raison, dès qu'on souhaite développer et maintenir une grammaire à large couverture, il est nécessaire de contrôler la grammaire complètement lexicalisée par une grammaire modulaire à partir de laquelle on la génère. Dans le cadre des TAG, il a été développé des formalismes modulaires à partir desquels on peut générer la grammaire TAG (Vijay-Shanker 1992, Candito 1996, 1999), ainsi que des procédures pour générer la grammaire TAG à partir d'une grammaire modulaire existante, comme HPSG (Kasper *et al.* 1995). Notre approche présente un avantage par le fait que nous proposons un formalisme qui permet d'écrire à la fois une grammaire modulaire et une grammaire complètement lexicalisée. On peut ainsi envisager de lexicaliser une partie de la grammaire seulement et de maintenir une grammaire non lexicalisée pour les constructions marginales.

J'aimerais insister, pour terminer cette section sur l'analyse verticale, sur le fait qu'une grammaire modulaire ne s'utilise pas nécessairement module par module. Quand on parle d'une architecture modulaire pour un système de TAL, on pense généralement, à tord, à une succession de modules agissant les uns après les autres. D'autre part, si j'ai mis l'accent sur le lien entre l'analyse verticale et les grammaires lexicalisées, c'est parce que ce lien existe et que les grammaires lexicalisées connaissent à l'heure actuelle un certain succès en TAL. Mais je ne voudrais pas que ceci masque le fait que l'analyse verticale est possible sans précompilation et qu'il s'agit, à mon avis, de la meilleure solution.

Enfin, tout ce que nous venons de montrer pour l'analyse est aussi valable pour la synthèse. La aussi, on peut envisager des stratégies horizontales ou verticales et il est possible d'utiliser la grammaire sous forme modulaire ou de la précompiler en une grammaire lexicalisée (cf. Danlos 1998, Candito & Kahane 1998 pour l'usage d'une grammaire complètement lexicalisée en synthèse).

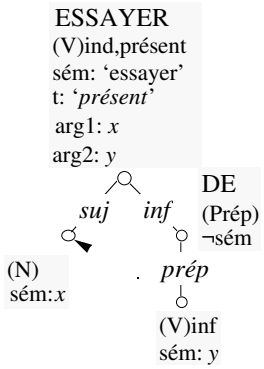

Figure 27 : Lexicalisation complète de *essayons*

Pour clore cette section sur les différentes stratégies dans la combinaison des règles, notons que les stratégies verticales et horizontales représentent les deux cas extrêmes et que des stratégies intermédiaires peuvent être envisagées. Par exemple, on peut envisager une stratégie verticale où les règles ne sont pas regroupées mots par mots, mais chunks par chunks.

5 Analyse en grammaire de dépendance

Après nos présentations des grammaires TST et GUST et de l'articulation des modules, nous allons nous concentrer sur le module qui pose les plus grandes difficultés en analyse, le module syntaxique de la grammaire, c'est-à-dire le module qui assure la correspondance entre une chaîne de mots et un arbre de dépendance. Nous considérerons les règles de syntaxiques présentées dans les Sections 0 et 4.2.2 (sous forme transductive, puis générative) : une telle règle associe une dépendance entre deux mots à une relation d'ordre entre les deux mêmes mots. Nous allons présenter trois techniques d'analyse : l'analyse par contrainte et l'analyse CKY, qui sont des stratégies d'analyse horizontale (les règles sont déclenchées module après module), et l'analyse incrémentale avec un analyseur à pile, qui est une stratégie verticale (les règles sont déclenchées mot après mot).

Nous illustrerons nos différentes techniques d'analyse sur l'exemple suivant :

(11) *Le boucher sale la tranche*

Cette phrase bien connue possède deux interprétations : 'le boucher est sale et il tranche quelque chose' ou 'le boucher met du sel sur la tranche'.

5.1 Analyse par contraintes

Le principe de l'analyse par contraintes est de considérer toutes les structures imaginables et de *filtrer* à l'aide des règles les structures bien formées qui peuvent correspondre à la phrase. Plutôt que tester l'une après l'autre toutes les structures imaginables (ce qui serait trop long), on construit en fait une structure très générale que l'on *contraint* par les règles et par les propriétés de bonne formation (par exemple le fait que l'on veuille un arbre projectif).

On commence donc par envisager pour chaque couple de mots de la phrase toutes les dépendances imaginables : on obtient ainsi un graphe de dépendance complet (Figure 28 de gauche). Ensuite, on applique les règles de linéarisation pour filtrer les dépendances qui sont validées par une règle de linéarisation (Figure 28 de droite)[38]. Rappelons qu'une règle de linéarisation dit, vu du point de vue de l'analyse, que si deux mots sont de telle et telle catégories et s'ils sont dans tel ordre, alors une dépendance avec telle fonction syntaxique peut les relier : par exemple, si un N suit un V, alors le N peut être l'objet du V.

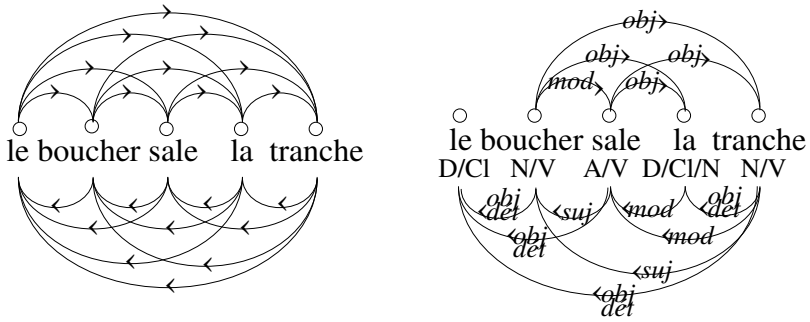

Figure 28 : Graphes de (11) avant filtrage et après filtrage par les règles de linéarisation

La dernière étape consiste à extraire des arbres projectifs du graphe ainsi obtenu (Figure 29). (Nous ne détaillons pas cette étape ; coir les sections suivantes pour cela).

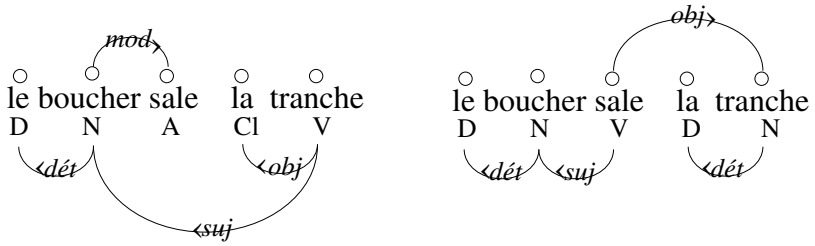

Figure 29 : Graphes de (11) après filtrage complet

L'analyse par contraintes est particulièrement adaptée aux grammaires de dépendances par le fait que, contrairement aux grammaires syntagmatiques, il est facile de considérer une structure qui contient en elle toutes les structures acceptables après filtrage. L'analyse par contraintes dans les

---

[38] Pour simplifier la présentation, nous utilisons des catégories lexicales très grossières. Par exemple, la catégorie Cl vaut pour tous les clitiques et comprend donc les (N,pro)nom et (N,pro)acc.

grammaires de dépendance a été introduite par Maruyama (1990a, 1990b) et développée, par exemple, par Duchier (1999 ; Duchier & Debusman 2001) ou par Blache (1998, 2001)[39].

Le même genre de techniques peuvent être appliquées avec des règles pondérées suivant leur probabilité d'apparition dans une situation donnée. Chaque règle possède un poids compris entre 0 et 1 ; plus le poids est proche de 0 plus la règle est contraignante. Après avoir construit le graphe de toutes les dépendances imaginables, on va utiliser les règles de linéarisation pour adresser à chaque dépendance un poids : soit le poids de la règle si une règle s'applique, soit le poids 0.1 si aucune règle ne s'applique (on évite les poids 0 qui écraseraient définitivement le score final). On pourra alors extraire du graphe l'arbre projectif qui donne le meilleur score (le *score* d'un arbre est le produit des scores des dépendances) (Menzel & Schröder 1998 ; Schröder *et al.* 2000). On pourra même accepter des entorses à la dépendance en pondérant également les règles qui assurent la projectivité. Plus généralement, pour des méthodes probabilistes en grammaire de dépendance, voir Eisner 1996, Collins 1997.

5.2 Analyse CKY

L'analyse CKY a été développée indépendamment par Cocke, Kasami et Younger pour les grammaires de réécriture hors-contextes (Kasami 1963, Younger 1967, Floyd & Biegel 1995). L'analyse CKY est une analyse montante : il s'agit d'identifier des segments analysables de la phrase de départ en allant des plus petits aux plus grands : si le plus grand segment analysable est la phrase complète, la phrase est donc analysable. L'algorithme fonctionne en temps $0(n^3)$ où n est le nombre de mots de la phrase. Avec une grammaire syntagmatique hors-contexte, on mémorise pour chaque segment analysé sa catégorie syntagmatique. L'algorithme peut être adapté trivialement aux grammaires de dépendance : dans ce cas, on mémorisera la catégorie lexicale de la tête du segment.

Considérons une phrase de longueur n. Pour chaque segment analysé allant du i-ième mot au j-ième mot (compris), on mémorise la catégorie X de la tête du segment sous la forme d'un triplet [i,j,X]. Avec le module morphologique, on commence par analyser tous les mots de la phrase, c'est-à-dire tous lessegments de longueur 1. Pour la phrase (11), on obtient :

[1,1,D], [1,1,Cl], [2,2,N], [2,2,V], …, [5,5,N], [5,5,V]

On essaye ensuite d'obtenir des segments de longueur 2 en utilisant les règles de linéarisation. Par exemple, pour la phrase (11), [1,1,D] + [2,2,N] = [1,2,N], car un élément de catégorie D à la gauche d'un élément de catégorie N peut dépendre de celui-ci par un dépendance *dét*. On obtient donc, pour la phrase (11) :

[1,2,N], [1,2,V], [2,3,N], [2,3,V], …, [4,5,N], [4,5,V]

Et ainsi de suite : les segments [i,j,X] et [j+1,k,Y] peuvent être combinés pour donner le segment [i,k,X] (resp. [i,k,Y]) s'il existe une règle de linéarisation indiquant qu'un élément de catégorie X précédant un élément de catégorie Y peut gouverner celui-ci (resp. peut dépendre de celui-ci). On construit ainsi tous les segments de longueur 2, 3, etc., jusqu'à n. Par exemple, pour construire les segments de longueur k (qui sont tous de la forme [i,i+k-*1*,Z]), on va considérer tous les couples de segments déjà obtenus de la forme ([i,j,X],[j+1,i+k-*1*,Y]) et chercher à les combiner par les règles de linéarisation. Ceci demande $k(n-k)C^2R$ opérations[40] où

---

[39] Blache 1998 considère au départ une grammaire syntagmatique avec tête à partir de laquelle il construit ensuite un graphe de dépendances.

[40] Pour k donné, on a n-k valeurs pour i, k valeurs pour j, C valeurs pour X et Y et R façons de combiner les segments à tester.

C est le nombre de catégories lexicales et R le nombre de règles de linéarisation. En sommant sur *k*, on obtient un résultat en $O(n^3C^2R)$.

Nous avons présenté l'algorithme de base. Tel quel cet algorithme vérifie que la phrase peut être associée à un arbre de dépendance projectif, mais il ne construit pas un tel arbre. Pour construire des arbres associés, le plus simple est de redescendre le calcul en partant des segments maximaux et de construire les arbres à partir de la racine. Pour l'exemple (11), le segment final [1,5,V] peut être obtenu de trois façons : en combinant [1,3,N] et [4,5,V] par la règle de placement du sujet, en combinant [1,3,V] et [4,5,N] par la règle de placement de l'objet ou en combinant [1,2,N] et [3,5,V] par la règle de placement du sujet. Comme les deux dernières correspondent au même arbre, on obtient en continuant les deux arbres de la Figure 29. On peut éviter de refaire les calculs en descendant l'arbre en conservant davantage d'informations lors du premier calcul (en indiquant pour chaque segment sa décomposition et la règle qui permet de l'obtenir), mais cela est en fait plus coûteux. Quoi qu'il en soit, il faut noter que le nombre d'arbres correspondant à une phrase de longueur *n* est dans le pire des cas une fonction exponentielle de *n* et que, par conséquent, un algorithme qui construirait *tous* les arbres de peut pas être polynomial (sauf à représenter la forêt d'arbres sous forme compacte).

Cet algorithme peut être enrichi de différentes façons.

1) Nous n'avons pas encore pris en compte le placement respectif des différents dépendants d'un même nœud. Celui-ci est encodé dans nos règles de linéarisation par les traits de *position* sur les dépendances. On peut très facilement tenir compte des positions en gardant en mémoire la dernière position utilisée pour chacune des deux directions : au lieu de segments [*i*,*j*,X], on manipulera des segments [*i*,*j*,X,*p*,*q*], où *p* est la dernière position utilisée pour un dépendant gauche de la tête du segment et *q* est la dernière position utilisée pour un dépendant droit (*p* et *q* étant égaux à 0 si aucune règle n'a été utilisé). Un tel segment ne pourra pas être combiné à un segment dépendant que par une règle dont le trait de position n'est pas compris entre *p* et *q* (le nouveau segment dépendant doit être positionné loin que les précédents). Pour l'exemple (11), si on combine les segments [4,4,Cl,0,0] et [5,5,V,0,0] avec la règle qui relie un clitique objet dans la position -4 au verbe, on obtiendra le segment [4,5,V,-4,0]. Ce segment ne pourra pas être combiné avec un clitique qui exige une position entre -4 et 0, mais pourra être combiné avec un élément qui accepte une position inférieure à -4.

2) Nous n'avons pas encore pris en compte les règles de sous-catégorisation, qui font partie des règles sémantiques de notre grammaire. On peut considérer cette information en indiquant dans la description d'un segment, en plus de la catégorie lexicale de la tête X, la liste des éléments sous-catégorisés par X qui ne sont pas dans le segment. Si on reprend l'exemple (11), le segment formé du seul mot *tranche*, lorsque ce dernier est analysé comme une forme du verbe TRANCHER, recevra une liste de sous-catégorisation avec *sujet* et *objet direct* : [5,5,V,{*suj*,*dobj*}]. Lorsque ce segment sera combiné avec le segment formé du mot *la* reconnu comme clitique accusatif, on obtiendra le segment [4,5,V,{*suj*}], par application de la règle de linéarisation du clitique *objet direct*. Lorsque, ce nouveau segment sera combiné avec le segment *le boucher sale* reconnu comme groupe nominal (et donc décrit comme [1,3,N,∅]), on obtiendra, par application de la règle de linéarisation du *sujet*, le segment [1,5,V, ∅]. A noter qu'on impose que, lors de la combinaison de deux segments, le segment dépendant soit saturé, c'est-à-dire que sa liste de sous-catégorisation soit vide. Quant à la liste de sous-catégorisation du segment tête, elle est privée de l'élément correspondant à la fonction du segment dépendant.

   Dans cet exemple, nous avons réduit l'information de niveau sémantique prise en compte au minimum. Si nous prenons en compte l'ensemble de l'information contenue dans la règle sémantique de la tête, notamment la description des dépendances sémantiques, nos descriptions de segments vont alors s'apparenter fortement aux descriptions de syntagmes en HPSG (Pollard & Sag 1994). Le mode de combinaison des segments que nous venons de décrire s'apparente lui-même au schéma de combinaison tête-actant d'HPSG (head-

daughter schema) : lorsque deux syntagmes X et Y se combinent pour en former un nouveau syntagme et que l'un des segments, par exemple Y, est reconnu comme un actant de la tête de l'autre, la description du nouveau segment est égale à la description du segment tête X où l'élément Y à été retiré de la liste des actants de X. La principale différence entre notre approche et HPSG est que la combinaison de deux segments doit être validée par une règle syntaxique séparée. En conclusion, en restant à un niveau de description grossier, on peut voir HPSG comme une version procédurale orientée vers l'analyse CKY d'une grammaire de dépendance.

3) Nous n'avons pas non plus pris en compte l'analyse des structures non projectives. Cela est possible. La modification de l'algorithme dépend de la façon dont sont écrites les règles qui assurent le placement des éléments qui ne sont pas dans la projection de leur gouverneur. Kahane *et al.* 1998 propose des règles de "lifting", permettant de remonter un élément sur un ancêtre de son gouverneur et de le positionner par rapport à ce dernier(cf. également Bröker 2000 pour une analyse commentée de cette solution). Dans la description d'un segment, on indiquera donc en plus de la catégorie lexicale de la tête et de sa liste de sous-catégorisation, la liste des éléments liftés. Encore une fois, cette solution s'apparente fortement aux descriptions de syntagmes en HPSG où apparaît un trait Slash (non-local) : le trait Slash contient précisément la liste des éléments "liftés", c'est-à-dire des éléments qui ne se placent pas dans la projection de leur tête, mais dans celle d'un ancêtre de leur tête.

   Avec les règles de "lifting", on peut encore obtenir un algorithme polynomial, mais il faut pour cela borner le nombre d'éléments "liftés" dans un segment (sinon le nombre de segments que l'on peut considérer croît exponentiellement avec n)..[41]

Remarquons que l'algorithme CKY est strictement montant et qu'une fois qu'un élément a été combiné avec son gouverneur il n'est plus possible de le combiner avec un de ses dépendants. (puisque le segment n'est représenté que par sa tête). Par exemple, si l'on analyse la phrase *Le garçon que j'ai rencontré la semaine dernière est étudiant*, il n'est pas possible de combiner *le garçon* avec la tête de la relative (*que* ou *ai* suivant les analyses) tant que la relative dans son entier n'a pas été analysée. La seule façon d'analyser le sujet de cette phrase est de combiner *la* et *dernière* à *semaine*, puis *la dernière semaine* à *rencontré*, puis le tout à *ai*, et ainsi de suite jusqu'à la combinaison de la relative complète avec *garçon*. Autrement dit, l'algorithme CKY, s'il a l'avantage d'être simple, ne peut en aucune façon être rendu incrémental.

Il existe un autre algorithme classique pour les grammaires hors-contextes, l'algorithme d'Earley (Earley 1970, Floyd & Biegel 1995), qui peut être aussi adapté aux grammaires de dépendance (Lombardo 1996). A l'inverse de l'algorithme CKY, l'algorithme d'Earley est un algorithme descendant. L'algorithme d'Earley fonctionne également en temps $0(n^3)$ (où n est le nombre de mots de la phrase) et même en temps $0(n^2)$ pour les grammaires non ambiguë. Néanmoins l'algorithme d'Earley n'est pas précisément adapté à la langue naturelle qui est hautement ambiguë. En particulier, cet algorithme, s'il apparaît comme plutôt incrémental, oblige en fait à construire l'arbre à partir de la racine et à anticiper, dès la lecture du premier mot, sur la chaîne complète de ces ancêtres dans l'arbre.[42]

---

[41] Il est probable qu'en l'absence d'une borne sur le nombre d'éléments liftés, le problème de la reconnaissance par une telle grammaire soit NP-complet (cf. Neuhaus & Bröker 1997 pour un résultat de ce type).

[42] Pour cette raison, l'algorithme d'Earley n'est pas applicable pour des grammaires décrivant des constructions récursives (un V peut subordonner un V qui subordonne un V qui …) sans introduire une limitation sur la profondeur de la récursion. Ceci est un vrai problème si on veut traiter de la langue naturelle. Par exemple, en français, anglais, allemand, etc., le degré d'enchâssement d'un groupe topicalisé en début de phrase est potentiellement infini.

5.3 Analyse incrémentale

On appelle *analyse incrémentale* une analyse qui se développe au fur et à mesure de la lecture.

Les différents algorithmes d'analyse incrémentale diffèrent sur le traitement de l'ambiguïté. Lorsque deux règles concurrentes sont applicables deux stratégies sont possibles : 1) choisir une des deux règles et en cas d'échec revenir en arrière (back-track) et essayer la deuxième règle ou 2) mener en parallèle les deux analyses. Nous appellerons *analyse incrémentale stricte* une analyse incrémentale qui ne permet pas de retour en arrière.

5.3.1 Analyse incrémentale et cognition

L'analyse incrémentale est la plus séduisante des techniques d'analyse du point de vue cognitif. Il est clair que les humains analysent un texte au fur et à mesure qu'il en prenne connaissance et qu'il peuvent parfaitement faire l'analyse d'un début de phrase et même en proposer des continuations.

Des expériences de psycholinguistique montrent par ailleurs que dans certains cas d'ambiguïté majeure, les sujets humains font des retours en arrière (O'Regan & Pynte 1992). Les exemples de ce type sont appelés des *garden paths*. Par exemple, lors de la lecture de (12b), on observe (par l'analyse du mouvement des yeux) au moment de la levée d'ambiguïté, lorsque *est* est considéré, une saccade régressive sur *reconduit*.

(12) a. *L'espion russe reconduit à la frontière un espion international.*
b. *L'espion russe reconduit à la frontière est un espion international.*

Dans la suite, nous allons donc nous intéresser à des algorithmes d'analyse incrémentale non stricte, lesquels correspondent davantage au fonctionnement humain. Nous reviendrons dans la Section 5.3.3 sur la question de savoir quelles sont les situations où doit être fait un choix, conduisant éventuellement à un échec, et quelles sont les situations où il faut éviter de traiter séparément deux options.

Du point de vue computationnel, il est permis de penser que dans la mesure où un système de TAL cherche à obtenir les mêmes résultats qu'un humain, la meilleure technique consiste à chercher à simuler au maximum la façon dont procède un humain.

5.3.2 Analyseurs à pile

Laissons de côté la question de l'ambiguïté pour le moment. On peut associer un analyseur à pile à un module syntaxique avec des règles de linéarisation comme celles que nous avons présentées dans les Sections 3.3.5 et 4.2.2 (qui indiquent quels sont les couples de mots qui peuvent être reliés entre eux). La technique consiste à charger dans la pile les mots au fur et à mesure de la lecture et à relier les mots par des opérations au sommet de la pile (Kornai & Tuza 1992, Kahane 2000b). Nous allons montrer sur un exemple comment fonctionne précisément l'analyseur à pile associé à notre module syntaxique. Nous commenterons nos règles sur l'exemple de la Figure 30.

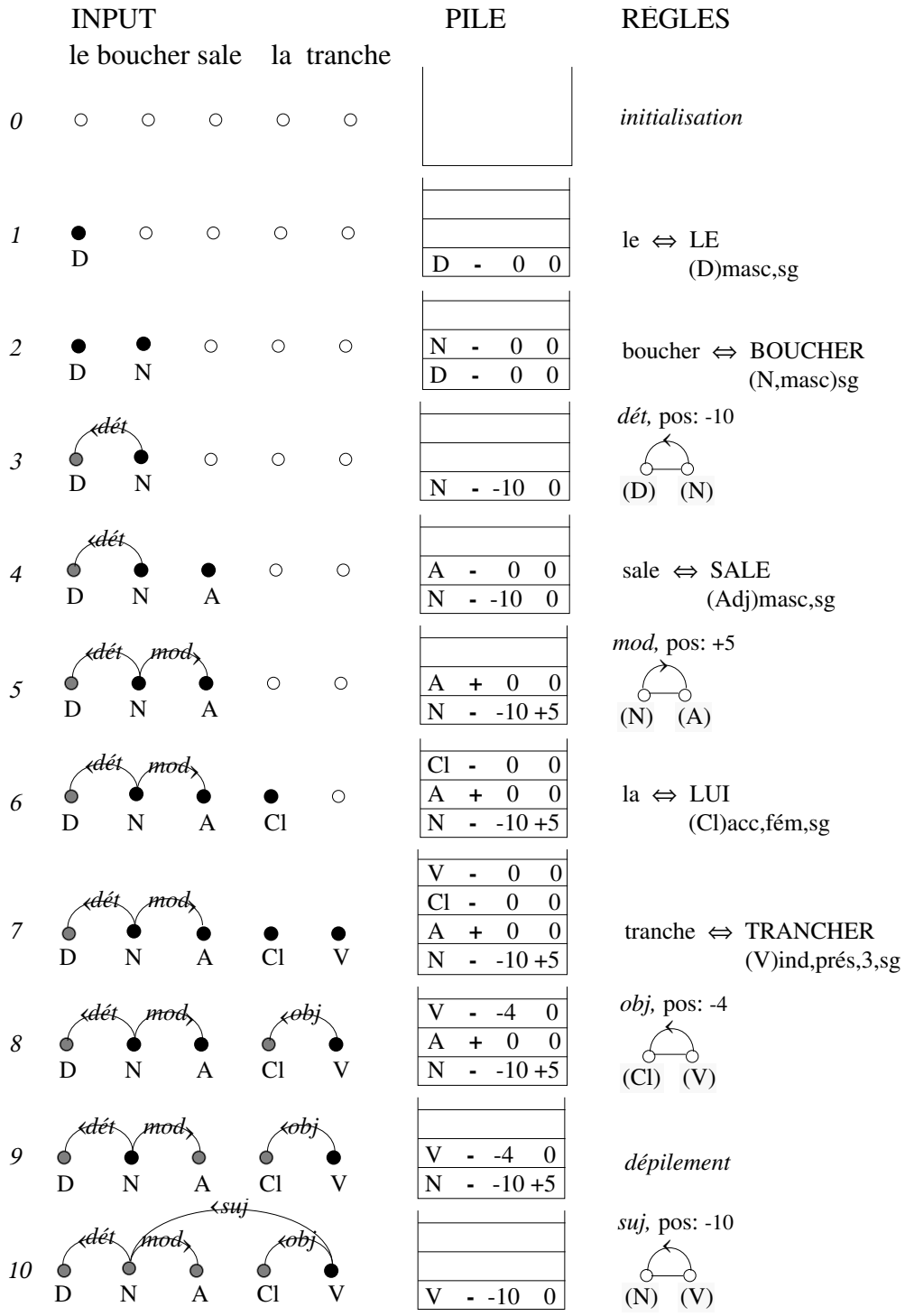

Figure 30 : Analyse de (11)

L'analyseur effectue la lecture de la phrase de gauche à droite. Au départ, la pile est vide (étape 0 : initialisation). Nous avons quatre types de règles de transition.

1) **Transition d'empilement** (étapes 1, 2, 4, 6 7). A chaque fois qu'un mot nouveau est lu, une règle morphologique est déclenchée et la catégorie du mot analysé est stockée dans la pile. Trois autres paramètres complètent la catégorie : le deuxième paramètre indique si le nœud est déjà gouverné ou non (**-** pour non gouverné, **+** pour gouverné), le troisième paramètre donne le valeur du trait de position de la dernière règle de linéarisation avec un

position négative utilisée et le quatrième paramètre donne la valeur du trait de position de la dernière règle de linéarisation avec une position positive utilisée. Lors du stockage, ces paramètres ont respectivement la valeur **-**, 0 et 0.

2) **Transition de liaison à un dépendant à gauche** (étapes 3, 8, 10). Une telle transition correspond à une règle de linéarisation dont le dépendant est à gauche. Les deux nœuds que nous allons lier se trouvent dans les deux cases supérieures de la pile. Appelons les *x* et *y*, *x* étant le dernier nœud lu et se trouvant sur le dessus de la pile. La règle peut s'appliquer si *x* et *y* possèdent les catégories requises par la règle de linéarisation. Le nœud *y* ne doit pas être déjà gouverné (valeur **-** du deuxième paramètre). Enfin, la valeur du trait de position de la règle doit être supérieure en valeur absolue à celle de la dernière règle avec une position négative utilisée pour relier *x* à un dépendant à sa gauche (voir le troisième paramètre). Ainsi à l'étape 8, le verbe *tranche* (= *x*) est lié avec le clitique *la* (= *y*) par un règle de position -4. Après application de la règle, le troisième paramètre de *tranche* prend la valeur -4. A l'étape 10, le verbe *tranche* est lié à son sujet par une règle de position -10, ce qui est possible car -10 est supérieur en valeur absolue à -4. Lors de l'application de la règle, la valeur -10 est consignée à la place de -4 dans la case de *tranche*. En raison de la projectivité, le nœud *y* est retiré de la pile. En effet, ce nœud ne peut avoir de dépendants à la droite de *x*, sans enfreindre la projectivité.

3) **Transition de liaison à un gouverneur à gauche** (étape 5). Une telle transition correspond à une règle de linéarisation dont le gouverneur est à gauche. Comme précédemment, les deux nœuds que nous allons lier se trouvent dans les deux cases supérieures de la pile. Appelons les *x* et *y*, *x* étant le dernier nœud lu et se trouvant sur le dessus de la pile. La règle peut s'appliquer si *x* et *y* possèdent les catégories requises par la règle de linéarisation. Le nœud *x* ne doit pas être déjà gouverné (valeur **-** du paramètre correspondant). Après la transition, les nœuds *x* et *y* sont tous les deux maintenus dans la pile, puisqu'ils peuvent avoir tous deux des dépendants à droite de *x*. Comme le nœud x est maintenant gouverné, la valeur du paramètre passe de **-** à **+**. Enfin, la valeur du trait de position de la règle doit être supérieure en valeur absolue à celle de la dernière règle de linéarisation avec une position positive utilisée pour relier *y* à un dépendant à sa gauche (voir le quatrième paramètre). A l'étape 5, le nom *boucher* (= *y*) n'a pas encore eu de dépendant à droite. Son quatrième paramètre est donc égal à 0. Après application de la règle, ce paramètre aura la valeur +5.

4) **Transition de dépilement** (étape 9). A tout moment, il est possible de retirer de la pile un nœud qui est déjà gouverné. Pour pouvoir lier *tranche* à son sujet *boucher*, on est ainsi obligé de dépiler l'adjectif *sale* (qui de toute façon, en raison de la projectivité ne pourra plus avoir de dépendant au -delà de *tranche*).

Une phrase est reconnue si à la fin de la lecture, la pile contient un unique nœud non gouverné, qui est en fait la racine de l'arbre de dépendance. Notons que nous assurons bien que le graphe construit est un arbre et que cet arbre est projectif.

L'analyseur en flux de Vergne (2000) utilise une méthode similaire à l'analyseur que nous venons de présenter, si ce n'est qu'il effectue un séquençage (chunking) préalable de la phrase et charge, au lieu des mots, les blocs (chunks) ainsi obtenus dans la pile (construisant ainsi un arbre de dépendance sur les blocs).

Comme pour l'analyseur CKY, cet analyseur peut être enrichi en prenant en compte des règles de plus haut niveau, notamment des règles de sous-catégorisation. On chargera alors dans la pile non seulement les caractéristiques morphologiques d'un mot (sa catégorie lexicale), mais aussi ses caractéristiques sémantiques. Voir Nasr 1995, 1996, Kahane 2000a, Lombardo 1992 pour des analyseurs de ce type (lesquels s'apparentent également aux grammaires catégorielles de Ajdukiewicz-Bar-Hillel). Une variante de ces analyseurs consiste à charger dans la pile non pas des mots, mais les liens potentiels qu'un mot peut avoir avec les mots qui le suivent. Cette méthode repose sur une description complète des valences possibles d'un mot. Le plus abouti

des analyseurs de ce type est la Link Grammar de Sleator & Temperley 1993. Dans ce formalisme lexicalisé, chaque règle décrit l'ensemble des liens que peut avoir un mot donné, c'est-à-dire les actants, mais aussi les modifieurs et les conjoints éventuels, ainsi que le gouverneur.

Ces méthodes peuvent être complexifiées pour traiter des arbres non projectifs : on peut, par exemple, garder en mémoire dans la pile dans la case d'un mot donné des informations sur certains de ses dépendants et autoriser les dits dépendant à créer des liens lorsque la case de leur gouverneur est considérée. Nous ne développerons pas cette question, par ailleurs fort intéressante, dans cette présentation (voir Nasr 1995, 1996, Kahane 2000a pour des traitements de ce type).

5.3.3 Traitement des ambiguïtés

Comment traiter les cas d'ambiguïté avec un analyseur incrémental ? La première technique consiste à faire des choix. A chaque fois que plusieurs règles concurrentes se présentent, il faudra choisir une règle. On peut développer des heuristiques, pour à chaque fois qu'un choix se présente, faire le "meilleur" choix. En cas d'échec, on peut effectuer un retour en arrière au dernier point où un choix a été fait et essayer le choix suivant. Si on autorise les retours en arrière sans mécanismes additionnels, on obtient un algorithme en temps exponentiel dans le pire des cas. En particulier, le pire des cas sera atteint à chaque fois qu'on aura affaire à une phrase agrammaticale (= pour laquelle notre grammaire ne peut fournir d'analyse). Pour obtenir un temps de traitement raisonnable, deux techniques sont possibles. La première consiste simplement à limiter les retours en arrière : on peut par exemple se fier entièrement aux heuristiques qui nous aident à faire le meilleur choix et interdire tout retour en arrière. On obtient alors un traitement en temps linéaire. C'est ce que fait l'analyseur en flux de Vergne 2000. Voir également Arnola 1998 pour un analyseur déterministe basé sur les dépendances. La deuxième technique, appeler mémoïsation, consiste, lors d'un retour en arrière, à conserver en mémoire les analyses déjà faites pour ne pas avoir à les refaire. Par exemple, si l'on considère l'exemple (11b) de garden-path, il ne sera pas nécessaire après le retour en arrière et le deuxième choix pour *reconduit* de refaire l'analyse de *à la frontière* (ce qui peut devenir vraiment intéressant si *frontière* gouverne en plus une relative) La mémoïsation, utilisée par Sleator & Temperley 1993 pour les Link Grammars, permet d'assurer une complexité en $0(n^3)$.

La deuxième technique de traitement des ambiguïtés consiste à ne pas faire de choix et à mener en parallèle les différentes analyses. Par exemple, Nasr 1995, 1996 utilise une technique adaptée de Tomita 1988 consistant à dupliquer la pile ; on peut ensuite factoriser un certain nombre d'opérations effectuées plusieurs fois dans plusieurs piles en factorisant les piles au sein d'une pile à structure de graphe et garantir un temps de traitement polynomial.

J'aimerais faire quelques commentaires sur la question du choix en me plaçant d'un point de vue linguistique et cognitif. Certaines grammaires, notamment des grammaires complètement lexicalisées comme TAG, considèrent des règles différentes pour chacune des sous-catégorisations d'un verbe (par exemple, *parler à Marie*, *parler de Jean*, *parler de Jean à Marie* correspondront à trois règles différentes pour *parler*). De telles grammaires obligent à faire des choix à tout va et ne font pas la différence entre des choix non pertinents (comme les différentes sous-catégorisation de *parler* qui devraient être traitées en parallèle) et des choix pertinents (comme les deux *reconduit* que l'on trouve en (11a) et (11b), qui diffèrent fortement puisque le premier gouverne le nom qui le précède, tandis que le deuxième en dépend). Je pense que la grammaire doit être écrite de telle façon que seuls les choix réels (les choix qui pourront conduire un locuteur à un retour en arrière s'il n'a pas fait le bon choix du premier coup) correspondent à des règles séparées. Par exemple, les différentes sous-catégorisations possibles d'un même verbe devront être rassemblées en une même règle. Il serait souhaitable d'évaluer précisément les situations qui provoquent des retours en arrière chez un locuteur afin de décider quand deux constructions doivent être traitées par une même règle et quand deux situations doivent correspondre à deux règles bien séparées entre lesquelles le locuteur doit faire un choix.

On pourra noter toute l'attention que nous avons portée à cette question dans l'écriture des règles de GUST (Section 4.2).

5.3.4 Limitation du flux

On peut observer sur les arbres de dépendances ordonnés des phrases d'une langue certaines limitations. Ainsi, bien qu'en l'absence de larges corpus étiquetés par des dépendances nous ne puissions être absolument affirmatif, il apparaît que le flux des dépendances est généralement borné par 6 ou 7 (voir Yngve 1960, Tuza & Kornai 1992 ou Murata *et al.* 2001 pour des hypothèses de cette nature). Nous appelons *flux des dépendances* en une position donnée (entre deux mots d'une phrase) le nombre de dépendances qui relient un mot à gauche de cette position à un mot à droite (Figure 31).

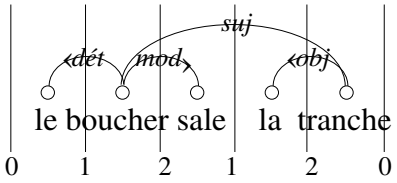

Figure 31 : Le flux des dépendances pour (11)

On peut penser que cette borne sur le flux correspond à une limitation liée à la mémoire immédiate: un locuteur ne peut gérer simultanément plus de 7 dépendances (voir la fameuse étude de Miller 1956 sur le fait que les humains ont au plus 7 ± 2 éléments dans leur mémoire à court terme).

Si l'on borne le flux des dépendances, on peut alors borner la taille de la pile dans l'analyseur à pile que nous avons présenté. Comme le langage de la pile est fini, le nombre de contenus possible de la pile est alors fini (bien que très gros). L'analyseur à pile, si l'on ne s'intéresse plus aux arbres de dépendance qu'il produit, est alors équivalent à un automate à nombre fini d'états (un état de l'automate est un contenu de la pile). Cet automate est donc équivalent à un automate déterministe, ce qui nous donne un reconnaisseur en temps linéaire (l'automate ne fournit plus d'analyse, mais peut seulement reconnaître les phrases qui ont une analyse). Cet automate peut être particulièrement utile pour filtrer les phrases agrammaticales, qui sont les phrases les plus coûteuses pour l'analyseur incrémental (puisque n'importe quel choix conduit à une situation d'échec et à un retour en arrière). Néanmoins, pour que cet automate ne soit pas trop gros, il faudra certainement limiter le nombre de symboles de pile (le nombre d'état de l'automate avant déterminisation est majoré par Z^k où Z est le nombre de symboles de pile et k le nombre maximum de nœuds autorisés dans la pile).

On peut également espérer optimiser l'analyseur incrémental par une étude statistique des contenus possibles de la pile pour des analyses correctes. Ceci permettrait dans une situation donnée de choisir entre des règles afin d'obtenir le contenu de pile le plus probable et d'éviter au maximum les échecs et les retours en arrière.

6 Conclusion

Comme nous l'avons dit au début de cet exposé, la dépendance est maintenant une notion utilisée par toutes les théories linguistiques, bien qu'elle soit souvent cachée sous diverses formes (fonctions syntaxiques, constituants avec tête, …). Nous espérons avoir convaincu le lecteur de l'intérêt qu'il y a à mettre en avant la dépendance et à écrire des règles qui manipulent explicitement des dépendances.

A travers l'étude des dépendances, nous avons souhaité mettre l'accent sur la théorie Sens-Texte. La TST est l'une des théories qui sépare le plus clairement les notions sémantiques, syntaxiques et morphologiques, en distinguant, en particulier, les dépendances sémantiques et syntaxiques, les lexies profondes et de surface, les grammèmes profonds et de surface ou en séparant clairement les règles de sous-catégorisation, d'ordre des mots, d'accord et de rection. D'autre part, la TST, en privilégiant la synthèse sur l'analyse, met bien en évidence l'avantage des grammaires de dépendance sur les grammaires syntagmatiques. En effet, la synthèse débute avec une représentation sémantique, à un moment où l'ordre des mots n'est pas encore fixé. Lorsqu'on veut décrire la synthèse d'une phrase en passant d'une représentation sémantique à une représentation où les mots sont ordonnés, on voit tout l'avantage qu'il y a à avoir un moyen de représenter la structure syntaxique sans avoir encore encoder l'ordre des mots ou même le regroupement des mots en constituants de surface. La synthèse met également l'accent sur l'importance des choix lexicaux et du lexique. En particulier, un notion comme celle de fonction lexicale prend toute son importance lorsqu'il faut faire les bons choix lexicaux (et ne pas dire *follement improbable* ou *hautement amoureux* à la place de *hautement improbable* ou de *follement amoureux*).

Nous avons également présenté une grammaire d'unification basée sur la TST, la Grammaire d'Unification Sens-Texte (GUST). Au delà de son intérêt propre, ce formalisme permet de rattacher plus facilement la TST à d'autres formalismes contemporains, comme HPSG, LFG, les Grammaires Catégorielles ou TAG. GUST hérite de la TST une claire séparation des informations sémantiques, syntaxiques et morphologiques et la modularité qui en résulte. Nous avons pu, au travers de GUST, montrer comment les règles de différents modules pouvaient être combinées, permettant, par exemple, d'écrire une grammaire complètement lexicalisée. D'autre part, contrairement aux grammaires complètement lexicalisées écrites dans d'autres formalismes (comme TAG), GUST permet de porter une grande attention à la façon dont les règles de la grammaire modulaire doivent être réparties entre les différentes lexies pour éviter une explosion du nombre de règles de la grammaire lexicalisée.

Nous avons terminé notre exposé par une présentation théorique des principales techniques d'analyse. Il faut noter que les grammaires de dépendances, à la différence des grammaires syntagmatiques, n'ont pas encore fait l'objet de travaux mathématiques ou d'informatique théorique d'envergure. Il n'existe pas pour les grammaires de dépendance de formalisme de référence (voir Kahane 2000b pour une proposition), comme le sont les grammaires de réécriture hors-contextes de Chomsky (1957) pour la grammaire syntagmatique. De même, tous les compilateurs de langages de programmation sont basés sur des techniques développées pour les grammaires hors-contextes. Nous espérons avoir montré que les mêmes techniques (comme l'algorithme CKY) se prêtaient au traitement des grammaires de dépendance et qu'en plus, les grammaires de dépendance permettaient des techniques propres, comme l'analyse incrémentale avec un analyseur à pile dont les cases de la pile contiennent les descriptions des mots de la phrase. D'autre part, nous avons pu faire le lien entre GUST et HPSG, montrant comment les grammaires syntagmatiques se présentent en fait comme des versions procédurales des grammaires de dépendance orientées vers l'analyse (et plus précisément l'analyse CKY, qui n'est pas, du point de vue cognitif et même computationnel, le plus intéressant des algorithmes d'analyse de la langue)[43].

Nous souhaiterions clore cet exposé, en évoquant ce que nous aurions aimé présenté et que nous n'avons pu présenter faute d'une maturité suffisante des notions concernées et d'un développement suffisant des travaux sur ces questions. Dans la Section 3.2.1, nous avons montré le rôle primordial que joue la structure communicative dans la représentation sémantique d'une phrase, mais nous n'avons pas pu montrer comment la structure communicative

---

[43] Quand on sait que les fondements de la grammaire syntagmatique reposent sur le distributionnalisme, c'est-à-dire sur une description des langues par la distribution des segments de textes, il n'est pas étonnant que la grammaire syntagmatique ait un lien étroit avec un algorithme de type CKY.

intervenait dans les différentes règles des différents niveaux. La structure communicative joue un rôle essentiel dans la hiérarchisation du graphe sémantique (notamment le choix de la tête syntaxique de la phrase) et dans la linéarisation. Dans les langues à ordre des mots relativement libre comme le russe ou l'allemand, la structure communicative (notamment la partition thème-rhème et la focalisation) contrôle fortement l'ordre des mots et la prosodie. Dans des langues à l'ordre moins libre, comme le français ou l'anglais, la structure communicative se réalise par des constructions particulières, comme le clivage, le pseudo-clivage ou la dislocation en français. D'autre part, nous n'avons pas abordé la question des constituants morphologiques : les mots, lorsqu'ils sont linéarisés, s'assemblent pour former des groupes qui sont placés les uns par rapport aux autres. Ces constituants morphologiques sont mis en évidence, entre autres, par la prosodie. La notion de constituant morphologique doit être distinguée de la notion de constituant syntaxique, laquelle n'est pas directement considérée en grammaire de dépendance.[44] Les constituants morphologiques forment une hiérarchie comparable aux constituants syntaxiques, mais il ne servent pas à représenter la structure syntaxique d'une phrase, laquelle est représentée, dans notre cadre théorique, par un arbre de dépendance. Parmi les constituants morphologiques, il faut en particulier distinguer les blocs (ou chunks) à l'intérieur desquels l'ordre des mots est très rigide et qui n'accepte pas de coupures prosodiques, comme les séquences déterminant-adjectifs-nom ou clitiques-verbe du français (Mel'čuk 1967, Abney 1991, Vergne 2000). Le rôle joué par de tels blocs (que ne considèrent d'ailleurs pas les grammaires syntagmatiques) n'est plus à faire en TAL, que ce soit pour l'analyse syntaxique ou la synthèse de la prosodie (Mertens 1997, Vergne 2000). La structure communicative joue un grand rôle, à coté de la structure de dépendance, dans la formation des constituants. Kahane & Gerdes 2001 propose, à partir de l'étude de l'ordre des mots en allemand, un formalisme qui permet d'associer à un arbre de dépendance une hiérarchie de constituants morphologiques, qui n'est pas le reflet immédiat de l'arbre de dépendance (et qui ne correspond donc pas non plus à une structure de constituants syntaxiques). Un même arbre de dépendance correspond à de nombreux ordres des mots et un même ordre des mots peut recevoir différentes structures de constituants morphologiques correspondant à différentes prosodies, mettant en évidence différentes structures communicatives. Ce travail doit maintenant être poursuivi pour montrer comment une structure communicative permet de choisir une structure de constituants morphologiques plutôt qu'une autre.

Références

Abeillé Anne, 1991, *Une grammaire lexicalisée d'Arbres Adjoints pour le français*, Thèse de doctorat, Université Paris 7, Paris.

Abeillé Anne, 1996-97, "Fonction objet ou position objet " (1ère et 2nde parties*), Gré des langues*, *11*, 8-29 ; *12*, 8-33.

Abney Steven, 1987, *The English Noun Phrase in its Sentential Aspect*, PhD thesis, MIT, Cambridge.

Abney Steven, 1991, "Parsing by chunks", in R. Berwick, S. Abney and C. Tenny (eds.), *Principle-Based Parsing*, Kluwer.

Abney Steven, 1992, "Prosodic structure, performance structure and phrase structure", *Proceedingsof Speech and Natural Language Workshop*, Morgan Kaufmann, San Mateo, CA.

---

[44] Comme nous l'avons montré dans la Section 2.1, les constituants syntaxiques considérés par les grammaires syntagmatiques peuvent être récupérés à partir de l'arbre de dépendance : ce sont les projections des sous-arbres constitués d'un nœud et de tout ou partie de ses dépendants. Par exemple, un constituant S ou Infl' est la projection d'un verbe, tandis qu'un constituant GV est la projection d'un verbe sans son sujet.

Anderson John, 1971, *The Grammar of Case: Towards a Localist Theory*, Cambridge University Press, Cambridge.

Ajdukiewicz Kasimir, 1935, "Die syntaktische Konnexität", *Studia Philosophica, 1*, 1-27.

Apresjan J., Boguslavskij I., Iomdin L., Lazurskij A., Sannikov V., Tsinman L., 1992, "ETAP-2: The linguistics of a machine-translation system", *Meta, 37:1*, 97-112.

Arnola Harri, 1998, "On parsing binary dependency structures deterministically in linear time", *Processing of Dependency-based Grammars, COLING/ACL'98 Workshop*, 68-77.

Bar-Hillel Yehoshua, 1953, "A quasi-arithmetical notation for syntactic description", *Language, 29.1*, 47-58.

Bar-Hillel Yehoshua, Gaifman Haïm, Shamir E., 1960, "On categorial and phrase-structure grammars", *Bull. Res. Counc. of Israël, 9F,* 1-16.

Blache Philippe, 1998, "Parsing ambiguous structures using controlled disjunctions and unary quasi-trees", *COLING/ACL'98*, Montréal, 124-30.

Blache Philippe, 2001, *Les grammaires de propriétés : des contraintes pour le traitement automatique des langues naturelles*, Hermès, 224p.

Blanche-Benveniste Claire, 1975, *Recherche en vue d'une théorie de la grammaire française. Essai d'application à la syntaxe des pronoms*, Champion, Paris.

Bloomfield Leonard, 1933, *Language*, New York.

Boyer Michel, Lapalme Guy, 1985, "Generating paraphrases from Meaning-Text semantic networks", *Computational Intelligence, 1*, 103-117.

Bresnan Joan (ed), 1982, *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge.

Bresnan Joan, Kaplan Ronald, Peters Stanley, Zaenen Annie, 1982, "Cross-serial dependencies in Dutch" *Linguistic Inquiry, 13:4*, 613-635.

Brody Michael, 1997, *Lexico-Logical Form: A Radically Minimalist Theory*, MIT Press, Cambridge.

Bröker Norbert, 2000, "Unordered and non-projective dependency grammars", *T.A.L., 41:1*, 245-272.

Candito Marie-Hélène, 1996, "A principle-based hierarchical representation of LTAG", *COLING'96*, Copenhagen.

Candito Marie-Hélène, 1999, *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées. Application au français et à l'italien*, Thèse de doctorat, Université Paris 7, Paris.

Candito Marie-Hélène, Kahane Sylvain, 1998, "Une grammaire TAG vue comme une grammaire Sens-Texte précompilée", *TALN'98*, Paris, 40-49.

Chomsky Noam, 1957, *Syntactic Structure*, MIT Press, Cambridge.

Chomsky Noam, 1965, *Aspects of the Theory of Syntax*, MIT Press, Cambridge.

Coch José, 1996, "Overview of AlethGen", *Proc. 8th Int. Workshop on Natural Language Generation (INLG'96), Vol. 2*, Herstmonceux, 25-28.

Coch José, 1998, "Interactive generation and knowledge administration in MultiMeteo", *Proc. 9th Int. Workshop on Natural Language Generation (INLG'98)*, Niagara-on-the-Lake, 300-303.

Courtin Jacques, Genthial Damien, "Parsing with dependency relations and robust parsing", *Workshop on Dependency-based Grammars, COLING/ACL'98*, Montréal, 25-28.

Danlos Laurence, 1998, "G-TAG : un formalisme lexicalisé pour la générationde textes inspiré de TAG", *T.A.L., 39:2*, 7-34.

Dikovsky Alexander, Modina Larissa, 2000, "Dependencies on the other side of the Curtain", *T.A.L.*, *41:1*, 79-111.

Duchier Denys, 1999, "Axiomatizing dependency parsing using set constraints", *Proc. 6th Meeting of the Mathematics of Language (MOL 6)*, Orlando, 115-126.

Duchier Denys, Ralph Debusmann, 2001, "Topological dependency trees: A constraint-based account of linear precedence", *ACL 2001*, Toulouse.

Dymetman Marc, Copperman Max, 1996, "Extended dependency structures and their formal interpretation", *COLING'96*, Copenhague, 255-61.

Earley J., 1970, "An efficient context-free parsing algorithm", *Communications of the ACM, 13:2*, 94-102.

Eisner Jason M., 1996, "Three new probabilistic models for dependency parsing: An exploration", *COLING'96*, Copenhague.

Engel Ulrich, 1992, *Deutsche Grammatik*.

Floyd Robert, Biegel Richard, 1995, *Le langage des machines : une introduction à la calculabilité et aux langages formels*, International Thomson Publishing, Paris.

Gaifman Haïm, 1965, "Dependency systems and phrase-structure systems", *Information and Control*, *18*, 304-337 ; Rand Corporation, 1961, RM-2315.

Garde Paul, "Ordre linéaire et dépendance syntaxique : contribution à une typologie", *Bull. Soc. Ling. Paris, 72:1*, 1-26.

Gerdes Kim, Kahane Sylvain, 2001, "Word order in German: A formal dependency grammar using a topological hierarchy", *ACL 2001*, Toulouse.

Gladkij Aleksej V., 1966, *Leckii po matematiceskoj linguistike dlja studentov NGU*, Novosibirsk (French transl: *Leçons de linguistique mathématique*, fasc. 1, 1970, Dunod).

Gladkij Aleksej V., 1968, "On describing the syntactic structure of a sentence" (en russe avec résumé en anglais), *Computational Linguistics, 7*, Budapest, 21-44.

Gross Maurice 1975, *Méthodes en syntaxe*, Hermann, Paris.

Hays David, 1960, "Grouping and dependency theories", Technical report RM-2646, Rand Corporation.

Hays David, 1964, "Dependency theory: A formalism and some observations", *Language*, *40:4*, 511-525.

Hellwig Peter, 1986, "Dependency Unification Grammar (DUG)", *COLING'86*, 195-98.

Hudson Richard, 1988, "Coordination and grammatical relations", *Journal of Linguistics, 24*, 303-342.

Hudson Richard, 1990, *English Word Grammar*, Oxford ,Blackwell.

Iordanskaja Lidija, 1963, "O nekotoryx svojstvax pravil'noj sintaksičeskoj struktury (na materiale russkogo jazyka)" [On some Properties of Correct Syntactic Structure (on the Basis of Russian)], *Voprosy Jazykoznanija, 4*, 102-12.

Iordanskaja L., Kim M., Kittredge R. I., Lavoie B., Polguère A., 1992, "Generation of extended bilingual statistical reports", *COLING'92*, Nantes, 1019-23.

Iordanskaja L., Mel'čuk I., 2000, "The notion of surface-syntactic relation revisited (Valence-controlled surface-syntactic relations in French)", in L.L. Iomdin, L.P. Krysin (ed), *Slovo v tekste i v slovare* [Les mots dans le texte et dans le dictionnaire], Jazuki Russkoj Kul'tury, Moscou, 391-433.

Jackendoff Ray, *X-bar Syntax. A Study of Phrase Structure*, MIT Press, Cambridge.

Jespersen Otto, 1924, *Philosophy of Grammar*, Londres.

Joshi Aravind, 1987, "Introduction to Tree Adjoining Grammar", in Manaster Ramer (ed), *The Mathematics of Language*, Benjamins, Amsterdam, 87-114.

Kahane Sylvain, 1996, "If HPSG were a dependency grammar ...", *TALN'96*, Marseille, 45-49.

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

Kahane Sylvain, 1998, "Le calcul des voix grammaticales", *Bull. Soc. Ling. de Paris*, *93:1*, 325-48.

Kahane Sylvain, 2000a, "Extractions dans une grammaire de dépendance lexicalisée à bulles", *T.A.L.*, *41:1*, 211-243.

Kahane Sylvain, 2000b, "Des grammaires formelles pour définir une correspondance", *TALN 2000*, Lausanne, 197-206.

Kahane Sylvain (ed), 2000c, *Grammaires de dépendance*, *T.A.L.*, *41:1*, Hermès.

Kahane Sylvain, 2001, "What is a natural language and how to describe it? Meaning-Text approaches in contrast with generative approaches", *Computational Linguistics and Intelligent Text Processing*, Springer, 1-17.

Kahane Sylvain, Mel'čuk Igor, 1999, "La synthèse sémantique ou la correspondance entre graphes sémantiques et arbres syntaxiques. Le cas des phrases à extraction en français contemporain", *T.A.L.*, *40:2*, 25-85.

Kahane Sylvain, Nasr Alexis, Rambow Owen, 1998, "Pseudo-projectivity: A polynomially parsable non-projective dependency grammar", *ACL/COLING'98*, Montréal, 646-52.

Kahane Sylvain, Polguère Alain (eds), 1998, *Workshop on Dependency-Based Grammars, ACL/COLING'98*, Montréal.

Kahane Sylvain, Polguère Igor, 2001, "Formal foundation of lexical functions", in B. Daille, G. Williams (eds), *Workshop on Collocation*, *ACL 2001*, Toulouse.

Kamp Hans, 1981, "Evènements, représentations discursives et référence temporelle", *Langages, 64*, 34-64.

Kamp Hans, Reyle Uwe, 1993, *From Discourse to Logic*, Kluwer, Dordrecht.

Kasami T., 1963, "An efficient recognition and syntax analysis algorithm for context-free languages," AFCRL-65-758, Air Force Cambridge Research Laboratory, Bedford, MA.

Kasper R., Kiefer B., Netter K., Vijay-Shanker K., 1995, "Compilation of HPSG to TAG", *ACL'95*.

Keenan Edward, Comrie Bernard, 1977, "Noun phrase accessibility and universal grammar", *Linguistic Inquiry, 8*, 63-100.

Kittredge Richard, Polguère Alain, 1991, "dependency grammars for bilingual text generation: Inside FoG's stratificational models", *Proc. Int. Conf. on Current Issues in Computational Linguistics*, Penang, 318-30.

Kornai Andreás, Tuza Zsolt, 1992, "narrowness, pathwidth, and their application in natural language processing", *Disc. Appl. Math, 36*, 87-92.

Lavoie Benoit, Rambow Owen, 1997, "RealPro: A fast, portable sentence realizer", *Proc. 5th Conf. On Applied Natural Language Processing (ANLP'97)*, Washington, 265-68.

Lecerf Yves, 1961, "Une représentation algébrique de la structure des phrases dans diverses langues natuelles", *C. R. Acad. Sc. Paris, 252*, 232-34.

Lecomte Alain, 1992, "Connection grammars: A graph-oriented interpretation", in Lecomte A. (ed), *Word Order in Categorial Grammar*, Adosa, Clermont-Ferrand, 129-48.

Lombardo Vincenzo, 1992, "Incremental dependency parsing", *ACL'92*, 291-93.

Lombardo Vincenzo, 1996, "An Earley-style parser for dependency grammars", *COLING'96*, Copenhague.

Lombardo Vincenzo, Leonardo Lesmo, 1998, "Formal aspects and parsing issues of dependency theory", *COLING/ACL'98*, Montréal, 787-93.

Lombardo Vincenzo, Lesmo Leonardo, 2000, "A formal theory of dependency syntax with empty units", *T.A.L., 41:1*, 179-210.

Maruyama Hiroshi, 1990a, *Constraint Dependency Grammar*, Technical Report RT0044, IBM, Tokyo.

Maruyama Hiroshi, 1990b, "structural disambiguisation with constraint propagation", *ACL'90*, Pittsburgh, 31-38.

Mel'čuk Igor, 1967, "Ordre des mots en synthèse automatique des textes russes", *T.A. Informations, 8:2*, 65-84.

Mel'čuk Igor, 1974, *Opyt teorii linguističeskix modelej "Smysl Tekst". Semantika, Sintaksis* [Esquisse d'une théorie des modèles linguistiques "Sens-Texte". Sémantique, Syntaxe], Moscou, Nauka, 314p.

Mel'čuk Igor, 1988a, *Dependency Syntax: Theory and Practice*, State Univ. of New York Press, Albany.

Mel'čuk Igor, 1988b, "Paraphrase et lexique: La Théorie Sens-Texteet le *Dictionnaire explicatif et combinatoire*", in Mel'čuk *et al*. 1988, 9-58.

Mel'čuk Igor, 1993-2001, *Cours de morphologie générale, Vol. 1-5*, Presses de l'Univ. de Montréal / CNRS.

Mel'čuk Igor, 1997, *Vers une Linguistique Sens-Texte*, Leçon inaugurale au Collège de France, Collège de France, Paris, 78p.

Mel'čuk Igor, 2001, *Communicative Organization in Natural Language (The Semantic-Communicative Structure of Sentences),* Benjamins, Amsterdam.

Mel'čuk Igor, Clas André, Polguère Alain, 1995, *Introduction à la lexicologie explicative et combinatoire*, Duculot, Paris.

Mel'čuk Igor, Pertsov Nikolaj, 1987, *Surface Syntax of English. A Formal Model within the Meaning-Text Framework*, Benjamins, Amsterdam.

Mel'čuk Igor, Žolkovsky Alexandr, 1984, *Explanatory Combinatorial Dictionary of Modern Russian*, Wiener Slawistischer Almanach, Vienne.

Mel'čuk Igor *et al.*, 1984, 1988, 1992, 1999*, Dictionnnaire explicatif et combinatoire du français contemporain, Vol. 1, 2, 3, 4*, Presses de l'Univ. de Montréal, Montréal.

Menzel Wolfgang, Schröder Ingo, 1998, 'Decision Procedures for Dependency Parsing Using Graded Constraints", *Workshop on Processing of Dependency-Based Grammars, COLING/ACL'98*, Montréal, 78-87.

Mertens Piet, 1997, "De la chaîne linéaire à la séquence de tons", *T.A.L., 38:1*, 27-52.

Milićević Jasmina, 2001, "A short guide to the Meaning-Text linguistic theory", in A. Gelbukh (ed), *Proc. of CICLing 2000*, à paraître chez Springer.

Miller George A., 1956, "The magical number seven, plus or minus two: Some limits on our capacity for processing information", *The Psychological Review, 63*, 81-97.

Murata Masaki, Uchimoto Kiyotaka, Ma Qing, Isahara Hitoshi, 2001, "Magical number seven plus or minus two: Syntactic structure recognition in Japanese and English sentences", in A. Gelbukh (ed), *Computational Linguistics and Intelligent Text Processing*, Springer, 43-52.

Nasr Alexis, 1995, "A formalism and a parser for lexicalised dependency grammars", *4th Int. Workshop on Parsing Tecnologies*, State Univ. of NY Press.

Nasr Alexis, 1996, *Un modèle de reformulation automatique fondé sur la Théorie Sens-Texte – Application aux langues contrôlées*, Thèse de doctorat, Université Paris 7, Paris.

Neuhaus Peter, Bröker Norbert, 1997, "The complexity of recognition of linguistically adequate dependency grammars", *ACL/EACL'97*, Madrid, 337-43.

O'Regan Kevin, Pynte Joël, 1992, "Regard et lecture", *Sciences cognitives, Courrier du CNRS n° 79*, CNRS, Paris, p. 16.

Owens Jonathan, 1988, *The Foundations of Grammar : An Introduction to Mediaeval Arabic Grammatical Theory*, Benjamins, Amsterdam.

PeŠkovskij Aleksandr, 1934, *Russkij sintaksis v naučnom osveščnii* [Syntaxe russe : une approche scientifique], Moscou, Učpedgiz.

Polguère Alain, 1990, *Structuration et mise en jeu procédurale d'un modèle linguistique déclaratif dans un cadre de génération de texte*, Thèse de doctorat, Université de Montréal.

Polguère Alain, 1992, "Remarques sur les réseaux sémantiques Sens-texte', in A. Clas (ed), *Le mot, les mots, les bons mots*, Presses de l'Univ. de Montréal, Montréal.

Polguère Alain, 1998, "Pour un modèle stratifié de la lexicalisation en génération de texte", *T.A.L., 39:2*, 57-76.

Pollard Carl, Sag Ivan, 1994, *Head-driven Phrase Structure Grammar*, Stanford CSLI.

Pustejovsky James, 1995, *The Generative Lexicon*, MIT Press, Cambridge.

Robinson Jane, 1970, "Dependency structures and transformational rules", *Language, 46*, 259-85.

Sag I., Gazdar G., Wasow T., Wisler S., 1985, "Coordination and how to distinguish categories", *Natural Language and Linguistic Theory, 3:2*, 117-171.

Schabes Yves, 1990, Mathematical and Computational Aspects of Lexicalized Grammars, PhD thesis, University of Pennsylvania, Philadelphie.

Schröder Ingo, Menzel Wolfgang, Foth Kilian, Schulz Michael, 2000, "Modeling dependency grammar with restricted constraints", *T.A.L.*, *41:1*, 113-44.

Schubert Klaus, 1987, *Metataxis: Contrastive Dependency Syntax for Machine Translation*, Foris, Dordrecht.

Sgall Petr, Hajicová Eva, Panenová Jarmila, 1986, *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Reidel, Dordrecht.

Sleator Daniel, Temperley Davy, 1993, "Parsing English with a Link Grammar", *Third Int. Workshop on Parsing Technologies* ; Carnegie Mellon Univ. Comp. Sc. Techn. Report CMU-CS-91-196, 1991.

Tesnière Lucien, 1934, "Comment construire une syntaxe", *Bulletin de la Faculté des Lettres de Strasbourg, 7*, 12ème année, 219-229.

Tesnière Lucien, 1959, *Éléments de syntaxe structurale*, Kincksieck, Paris.

Tomita Masaru, 1988, "Graph structured stack and natural language parsing", *ACL'88*, Buffalo.

Vergne Jacques, 2000, *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur - Analyse syntaxique automatique non combinatoire*, Thèse d'HDR, Université de Caen.

Vijay-Shanker K., Yves Schabes, 1992, "Structure sharing in Lexicalized TAG", *COLING'92*.

Wanner Leo (ed), 1996, *Lexical Functions in Lexicography and Natural Language Processing*, Benjamins, Amsterdam.

Weiss Daniel, 1999, "Sowjetische Sprachmodelle und ihre Weiter fürhung", *Handbuch des sprachwissenschaftlich Russistik und ihrer Grenzdisziplinen*, Harrassowitz, 973-09.

XTAG Research Group, 1995, "A Lexicalized Tree Adjoining Grammar for English", technical Report IRCS 95-03, University of Pennsylvania (version mise à jour sur le web).

Yngve Victor H., 1960, "A model and an hypothesis for language structure", *The American Philosophical Society, 104:5*, 444-66.

Yngve Victor H., 1961, "The Depth Hypothesis", *Proceedings of Symposia in Applied Mathematics, Vol. 12: Structure of Language and its Mathematical Aspects*, American Mathematical Society, Providence, 130-138.

Younger D.H., 1967, Recognition of context-free languages in time n^3", *Information and Control, 10:2*, 189-208.

Žolkovskij Aleksandr, Mel'čuk Igor, 1965, "O vozmožnom metode i instrumentax semantičeskogo sinteza" [Sur une méthode possible et des outils pour la synthèse sémantique (de textes)], *Naučno-texničeskaja informacija [Scientific and Technological Information],* 6, 23-28.

Žolkovskij Aleksandr, Mel'čuk Igor, 1967, "O semantičeskom sinteze" [Sur la synthèse sémantique (de textes)], *Problemy Kybernetiki [Problèmes de Cybernétique], 19,* 177-238. [trad. franç. : 1970, *T.A. Information, 2*, 1-85.]

Zwicky Arnold, 1985, "Heads", *Journal of Linguistics, 21*, 1-29.