Grammaire d'Unification Sens-Texte
Vers un modèle mathématique articulé de la langue

Document de synthèse pour l'habilitation à diriger les recherches

Sylvain Kahane

1 Introduction

> Plus on aime le neuf, plus il faut se soucier de garder intact et fort le legs du passé qui va permettre au neuf de s'établir et lui donner le temps de trouver sa forme.
>
> Lyautey (1854-1934)

Ce mémoire présente un ensemble de travaux qui m'amène à proposer une nouvelle architecture pour la modélisation mathématique des langues naturelles. Le modèle qui en résulte, que j'appelle la *Grammaire d'Unification Sens-Texte* [dorénavant abrégé en *GUST*], repose sur une synthèse de divers courants en modélisation des langues actuellement très actifs : HPSG (*Head-driven Phrase Structure Grammar*), TAG (*Tree Adjoining Grammar*), LFG (*Lexical Functional Grammar*), les grammaires catégorielles et, avant tout, les grammaires de dépendance et la théorie Sens-Texte [dorénavant TST]. Comme nous le verrons dans les Sections 1.2 à 1.5, chacun des modèles considérés posent des problèmes spécifiques que GUST tente de résoudre.

1.1 A propos du foisonnement des modélisations

Immanquablement, une telle étude pose la question de savoir s'il est bien raisonnable de proposer encore une nouvelle famille de modèles alors qu'il y en déjà tant.

Il est vrai que ce foisonnement de modèles et de formalismes est un handicap pour la linguistique et tout particulièrement pour les ceux qui veulent décrire une langue particulière, pour qui le choix d'un cadre de modélisation plutôt qu'un autre est un choix difficile dont ils ne maîtrisent pas forcément les tenants et les aboutissants. Le résultat en est que des linguistes qui travaillent sur les mêmes langues ou les mêmes phénomènes en arrivent à ne pas se comprendre ou à s'opposer sans voir que leurs conclusions sont similaires.

Ce foisonnement est aussi le signe qu'aucun modèle linguistique n'est suffisamment satisfaisant pour s'être imposé.[1] Il a permis, et permet toujours, à de nouvelles idées de se développer plus facilement. Il entraîne la multiplication des descriptions d'un même phénomène, permettant ainsi d'en mettre à jour toutes les facettes, d'en mieux comprendre les mécanismes et de tendre vers la description la plus juste.

Mais surtout ce foisonnement n'est pas aussi foisonnant qu'on pourrait le penser au premier abord. Les différents modèles auxquels nous ferons référence ici montrent, malgré des différences d'architecture parfois importantes, une remarquable convergence théorique. En particulier, tous ces modèles :

- distinguent plusieurs niveaux de représentation et distinguent notamment, plus ou moins clairement, les *dépendances sémantiques* (= relations prédicat-argument)[2] et les *dépendances syntaxiques* (= relations fonctionnelles = relations tête-fille)[3] et parfois les *constituants de surface* (= constituants topologiques)[4] ;

---

[1] La situation a largement évolué par rapport aux années 70 et à l'omniprésence de la grammaire générative-transformationnelle. Néanmoins, le courant chomskyen, avec aujourd'hui le Programme Minimaliste, reste le courant majoritaire. Je ne me situerai pas directement par rapport à ce courant, mais par contre je considérerai ici plusieurs modèles qui eux mêmes se situent par rapport au courant chomskyen, comme LFG ou HPSG. Joan Bresnan et Ivan Sag, qui sont respectivement à l'origine des modèles LFG et HPSG et qui ont fait leur thèse au MIT et ont débuté dans la grammaire générative-transformationnelle, ont amplement justifié des raisons qui les ont amenés à s'en détacher (cf. Bresnan 1982 et Pollard & Sag 1994).

[2] La distinction entre dépendances sémantiques et syntaxiques apparaît dès les premiers travaux en TST (Mel'čuk & Žolkovskij 1967). Voici les exemples usuels (voir aussi [D1]) :

(i) a. *Pierre lisse la surface*
b. *une surface lisse*
c. *La surface semble lisse*

Dans ces trois exemples, le sens du mot *lisse* se comporte comme un prédicat (sémantique) qui prend pour argument le sens du mot *surface*. Dans l'exemple (i-a), la dépendance sémantique double une dépendance syntaxique de même sens (*surface* est un actant de *lisse*). Dans l'exemple (i-b), la dépendance syntaxique entre *lisse* et *surface* est de sens inverse : *surface* est à la fois le gouverneur syntaxique et l'argument sémantique de *lisse* (*lisse* est un modifieur de *surface*). Dans l'exemple (i-c), *surface* est toujours l'argument sémantique de *lisse*, mais il n'y a plus de dépendance syntaxique entre *lisse* et *surface* ; inversement, *surface* est un dépendant syntaxique de *semble* (son sujet syntaxique pour être précis), mais il n'y a pas de dépendance sémantique entre *surface* et *semble* ; le seul argument sémantique de *semble* est *lisse*, comme dans *il semble que la surface est lisse*.

[3] Dans la Syntaxe X-barre (Jackendoff 1977) et toutes les grammaires syntagmatiques qui en découlent, les dépendances syntaxiques ne sont pas explicites, mais sont néanmoins considérées par le fait que chaque constituant est la projection d'une tête lexicale. Voir [C2] pour le passage d'une structure syntagmatique avec têtes à une structure de dépendance et vice-versa.

[4] Par constituants de surface ou constituants topologiques, nous entendons les constituants qui interviennent dans le calcul de l'ordre des mots et des groupes prosodiques, comme les *chunks* (Abney 1991) :

(i) (*je parle*) (*à la sœur*) (*du lieutenant*)

Ces constituants n'ont pas de rapport direct avec les constituants des grammaires basées sur la Syntaxe X-barre.

- font une large place au lexique et déduisent la description d'une phrase de la combinaison des descriptions des unités lexicales de cette phrase, ce qui est connu sous le terme de *grammaires lexicalisées*.

Malgré cela, l'architecture du lexique reste insuffisante dans tous ces modèles et le traitement des morphèmes non lexicaux n'est pas aussi propre que celui des morphèmes lexicaux, comme nous allons le montrer dans les sections qui suivent.

Avant de présenter les grandes lignes de GUST, j'aimerais inventorier rapidement les avantages et les problèmes des différents modèles par rapport auxquels GUST se situe.

1.2 La théorie Sens-Texte (TST)

La principale source d'inspiration de GUST est la TST (Mel'čuk 1988, 1997, [A4], [D1] : Section 3), dont GUST reprend les principaux choix théoriques.

GUST repose sur les postulats centraux de la TST :

1) un modèle linguistique doit avant tout modéliser la correspondance entre les sens et les textes d'une langue (*texte* = production graphique ou phonique) ;

2) la correspondance est modulaire et passe par au moins deux niveaux de représentation intermédiaires clairement identifiés : un niveau syntaxique (qui rend compte de la structure de la phrase, c'est-à-dire de l'organisation des mots dans la phrase) et un niveau morphologique (qui rend compte de la structure interne des mots).

En fait, la TST considère 7 niveaux de représentation et introduit en plus des 4 niveaux standard — sémantique, syntaxique (de surface), morphologique (de surface) et phonologique (de surface) — 3 niveaux intermédiaires : syntaxique profond, morphologique profond et phonologique profond.

niveau sémantique
niveau syntaxique profond
niveau syntaxique de surface
niveau morphologique profond
niveau morphologique de surface
niveau phonologique profond
niveau phonologique de surface

La raison pour laquelle la TST considère des niveaux intermédiaires est que le passage d'un niveau standard à l'autre, implique deux changements majeurs : un changement de structure[5] et un changement d'unités de base. Ainsi le passage du niveau sémantique au niveau syntaxique (de surface) met-il en jeu la hiérarchisation (le passage d'un graphe à un arbre) et la lexicalisation (le passage des sémantèmes aux mots). Les représentations intermédiaires sont introduites dans le but de séparer les deux opérations ; ainsi, la représentation syntaxique profonde reprend-elle a peu près les mêmes unités que la représentation sémantique et la même structure que la représentation syntaxique (de surface). Je pense pour ma part qu'on ne gagne rien à considérer de telles représentations intermédiaires et qu'on ne peut séparer la hiérarchisation de la lexicalisation. GUST considère donc uniquement les 4 niveaux standard. Néanmoins, comme nous le verrons dans la Section 3.3, les 3 représentations intermédiaires de la TST, bien que n'étant pas considérées directement, sont sous-jacentes à l'architecture de GUST.

Il est important de souligner que la représentation dite sémantique de la TST (reprise par GUST) ne se situe pas à un niveau aussi profond que les représentations généralement manipulées par les sémanticiens ; l'objectif est ici de répertorier les sens exprimés dans la phrase par les signes lexicaux et grammaticaux et d'indiquer les dépendances sémantiques (ou relations prédicat-argument) entre ces sens ; il n'est pas d'indiquer l'état du monde dénoté par le sens de la phrase et de permettre des inférences logiques. Sur le graphe que forment les dépendances sémantiques se superposent d'autres structures dont nous ne discuterons pas ici et qui indiquent, entre autres, l'organisation communicative (angl. *information structure*) ou la structure référentielle. De même que le squelette de la représentation sémantique TST est un graphe, le squelette de la représentation syntaxique de surface est un arbre de dépendance entre les mots de la phrase (Tesnière 1959). Dans mes travaux sur l'extraction (et aussi sur la coordination), sur lesquels je reviendrai, j'ai proposé de considérer des structures syntaxiques plus riches, les arbres à bulles ([C2], [A5], [A3]). Le squelette de la représentation morphologique est la suite des représentations morphologique des mots. Alors que la majorité des théories donnent à la constituance un rôle fondamental, la TST considère uniquement les constituants comme une structure annexe de la représentation morphologique[6] et ceux-ci n'avaient jamais fait l'objet d'étude dans le cadre de la TST jusqu'à aujourd'hui. Ce manque a été en partie comblé par mes travaux avec Kim Gerdes sur l'ordre des mots en allemand ([C16], [C18]) (voir aussi Milićević *à paraître* pour un travail sur l'ordre des mots en serbe

[5] J'emploie ici le terme *structure* au sens mathématique pour désigner le type d'organisation structurelle de la représentation : graphe, arbre, ordre linéaire, etc.

[6] D'autres traditions en grammaire de dépendance (comme Sgall *et al*. 1986 ou Hudson 1990, 2000) rejettent complètement la constituance. En fait, les constituants syntaxiques tels que considérées par la syntaxe X-barre (Jackendoff 1977) sont redondants avec une structure de dépendance (voir par exemple [C2] ou [D1]). Les constituants considérés dans le cadre de la TST sont évidemment des constituants de surface, qui n'ont pas pour objectif d'encoder les relations de dépendance syntaxiques, mais plutôt les regroupements observés en surface qui dépendent tout autant de la structure communicative que de la structure de dépendance et qui s'interfacent assez naturellement avec les constituants prosodiques.

allant dans la même direction). Nous avons également débuté avec Igor Mel'čuk un travail sur la caractérisation théorique des constituants minimaux en français (les *chunks*).

Le principal défaut de la TST, outre la trop grande simplicité structurelle des représentations considérés, est le manque de formalisation du modèle et l'absence d'un modèle pour une langue donnée qui soit suffisamment développé pour servir de modèle de référence. Quelques efforts de formalisation ont été fait, notamment autour de projets d'implémentation de la TST pour des générateurs de texte (Iordanskaja, Kittredge & Polguère 1988, Bohnet & Wanner 2001). D'autres propositions de modélisations ont été faites qui sont basées sur les principes généraux de la TST sans en garder les mécanismes exacts, notamment les règles de correspondance (Boyer & Lapalme 1985, Nasr 1995, 1996). L'article [A2], écrit avec Igor Mel'čuk, est probablement la première présentation entièrement formalisée de l'interface sémantique-syntaxe[7] de la TST, c'est-à-dire du module qui assure la correspondance entre graphes de dépendances sémantiques et arbres de dépendance syntaxiques. La complexité des phénomènes considérés — la relativisation et l'interrogation (indirecte) — nous a obligés à développer des aspects jusque-là peu étudiés de la TST comme le formalisme des règles sémantico-syntaxiques, l'organisation générale de l'ensemble des règles, le rôle de la structure communicative dans le passage à une structure syntaxique hiérarchisée et surtout l'aspect procédural de la synthèse, c'est-à-dire la mise en œuvre du passage d'un graphe sémantique à un arbre de dépendance syntaxique. (Il faut savoir que la présence de phénomènes d'extraction entraîne de fortes discontinuités entre les deux niveaux de représentation envisagés et rend par conséquent le passage particulièrement délicat.) Nous avons mis en évidence l'insuffisance des précédentes formalisations qui n'indiquaient pas assez explicitement dans les règles quelles étaient les parties de la structure réellement manipulées par la règle en question et quelles étaient les parties de la structure qui faisaient office de contraintes ou de pointeurs vers d'autres règles.

Enfin, on peut considérer que GUST, poursuivant en cela les travaux de Nasr 1995, 1996, est en soi une réponse à la formalisation de la TST. Néanmoins, GUST n'est pas une formalisation "kascher" de la TST pour plusieurs raisons sur lesquelles nous aurons l'occasion de revenir.

1.3 La grammaire d'adjonction d'arbres (TAG)

La deuxième source d'inspiration de GUST est certainement TAG (Joshi, Levy & Takahashi 1975, Joshi 1987, Abeillé 1991).

Comme TAG, GUST repose sur l'idée que la structure d'une phrase est obtenue par la combinaison de structures dites élémentaires associées aux unités de la phrase. L'idée vient en fait des grammaires catégorielles (Ajdukiewicz 1935, Bar-Hillel 1953) : contrôler la

[7] Dans la plupart de mes publications liées à la TST, j'utilise la terminologie usuelle en TST de module sémantique pour l'interface sémantique-syntaxe, module syntaxique pour l'interface syntaxe-morphologie et module morphologique pour l'interface morphologie-phononologie.

combinatoire des mots de la langue par des descriptions associées aux mots et des opérations de combinaisons de ces descriptions. Le modèle d'une langue se réduit alors à un lexique, lequel contient toute la grammaire (et les modèles linguistiques de ce type sont appelés des *grammaires complètement lexicalisées*). TAG se différencie des grammaires catégorielles par le fait que les descriptions associées aux mots sont des structures géométriques — des portions d'arbre — et que la combinaison des descriptions des mots, en plus de contrôler la combinatoire des mots, fournit directement une structure pour la phrase résultante. Les tenants de la grammaire catégorielle se sont également attachés à structurer davantage les descriptions attachées aux mots, notamment avec l'émergence des grammaires catégorielles logiques (Lambek 1958, Moortgat 1988, Morryll 1994), où les descriptions attachées aux mots deviennent des formules logiques et la dérivation d'une phrase s'apparente à une preuve logique. Enfin, des travaux sur la géométrie des preuves mettent finalement en évidence une certaine parenté entre grammaires lexicalisées géométriques à la TAG et grammaires catégorielles logiques (Lecomte 1992, Rétoré 2001). Ma préférence va néanmoins à des formalismes comme TAG qui explicitent les structures géométriques qu'ils manipulent, même si cela n'a pas forcément de conséquences sur l'adéquation descriptive des modèles.

Comme nous le verrons, GUST n'est pas une grammaire complètement lexicalisée. Elle se distingue de TAG (et des grammaires catégorielles) par le fait que les structures élémentaires sont associées à différents types de signes, y compris les signes grammaticaux, et que la structure associée à un mot n'est pas élémentaire, mais est obtenue par la combinaison de plusieurs structures associées à différents signes appartenant à différents modules.

Une des principales qualités de TAG, que GUST a tenté de préserver, est la grande simplicité du formalisme. Seules deux opérations de combinaison des structures élémentaires sont possibles en TAG : la substitution, reprise des grammaires de réécriture hors-contexte, et l'adjonction, qui permet un traitement élégant des modifieurs et des dépendances non bornées. A toute dérivation d'une phrase par une grammaire TAG (c'est-à-dire à toute combinaison de structures élémentaires donnant une phrase), on peut associer un arbre de dérivation dont chaque nœud renvoie à une unité lexicale de la phrase et à la structure élémentaire choisie pour la représenter et chaque branche indique le type de combinaison entre les deux structures élémentaires qu'elle relie. Avec Marie-Hélène Candito, nous avons montré dans [C6] que, en respectant certains principes linguistiques généralement considérés dans l'écriture des grammaires TAG (mais pas toujours appliqués, ni applicables), un arbre de dérivation TAG pouvait être interprété comme un graphe de dépendances sémantiques. Ce travail a été poursuivi par d'autres travaux sur l'interface sémantique-syntaxe en TAG, notamment Joshi & Kallmeyer 1999.

Je souhaiterais maintenant insister sur deux des problèmes principaux que pose TAG et que GUST tente de pallier.

Premièrement, TAG est peu articulé et effectue directement la correspondance entre le niveau morphologique et le niveau sémantique. Il en résulte diverses difficultés : pas de traitement des signes grammaticaux en tant que tels (ceux-ci ne reçoivent pas de structure propre), un

traitement coûteux des locutions (celles-ci reçoivent une structure complète sans lien avec les structures canoniques de leurs composantes), la concurrence entre arguments sémantiques et syntaxiques (par exemple, dans la construction *Pierre semble malade*, la combinaison entre *semble* et son sujet syntaxique *Pierre* est évacuée au bénéfice de la combinaison entre *malade* et son argument sémantique *Pierre*), un traitement difficile des langues à ordre libre, etc. Concernant le dernier point, les travaux de Kim Gerdes (*en préparation*), à la suite de ceux d'Owen Rambow (1994), montre que si l'on veut maintenir des principes linguistiques forts (permettant, par exemple, d'interpréter l'arbre de dérivation comme le squelette d'une représentation sémantique, suivant [C6]), TAG ne permet de traiter qu'un fragment de l'allemand (en raison de phénomènes comme l'extraposition des relatives ou le *partial VP fronting*).

Le deuxième problème posé par TAG est l'explosion combinatoire du nombre de structures élémentaires associées à un même lexème ou à un même mot-forme. Par exemple, la grammaire FTAG du français développée à Talana[8] (Abeillé 1991, Candito 1999, Abeillé & Candito 2001) possède plus de 5000 structures élémentaires pour les verbes, sans pour autant traiter des phénomènes aussi répandus que la négation en *ne … personne/rien* (*Pierre ne veut rien manger*), les causatifs en *faire* Vinf (*Pierre le fera parler*), la dislocation (*Pierre, je veux lui parler*) ou la coordination (*Pierre mange et parle à Marie en même temps*). La raison de cette explosion combinatoire, soulignée dans [C10], vient du formalisme qui oblige à spécifier dans la structure élémentaire d'un verbe la position de ses arguments syntaxiques/sémantiques. Ainsi, alors qu'il serait plus économique pour dériver *Pierre le regarde* d'utiliser la structure canonique de *regarde* et d'indiquer dans la structure associée au clitique *le* le placement particulier de ce pronom (voir le traitement proposé en GUST dans [A3]), la grammaire TAG considère au contraire des structures différentes pour *regarde* avec un objet canonique (*Pierre regarde un cochon*) et *regarde* avec un objet clitique (*Pierre le regarde*) ou encore *regarde* avec un objet extrait (*Je me demande quel cochon Pierre regarde*). Ainsi pour chaque phénomène considéré, une structure différente est nécessaire et l'explosion combinatoire naît simplement du croisement des phénomènes.

Cette explosion combinatoire est encore accentuée par certains choix linguistiques. On peut, en restant dans le cadre standard des grammaires TAG, réduire un peu le nombre actuel de structures élémentaires (en améliorant de surcroît la couverture et en simplifiant l'interface sémantique-syntaxe). Dans [C9], nous montrons que le fait qu'une relative puisse modifier un nom peut être associé au mot *qu-* plutôt qu'au verbe gouvernant le mot *qu-*. Ceci permet de considérer la même structure élémentaire pour tous les types d'extraction (topicalisation, relativisation, interrogation, etc.). Cette analyse n'a pas été implantée dans les grammaires TAG du français ou de l'anglais et l'architecture de ces grammaires rendrait de telles modifications assez périlleuses.

[8] La jeune équipe Talana, dirigée par Laurence Danlos et à laquelle j'appartiens depuis 1996, s'est fondue en janvier 2001 dans l'équipe Lattice (UMR 8094), dirigée par Catherine Fuchs.

La couverture et l'explosion combinatoire ne sont pas les seuls problèmes des grammaires TAG. Une grammaire de 5000 arbres ne peut pas être écrite et maintenue sans outils informatiques. La version actuelle de FTAG a été regénérée automatiquement (et corrigée et étendue) à partir d'une surcouche par Marie-Hélène Candito dans le cadre de sa thèse (Candito 1999). Cette surcouche, appelée la métagrammaire, inventorie les différents diathèses possibles et les différentes réalisations possibles de chaque actant et permet de générer automatiquement l'ensemble des structures élémentaires possibles par croisement des différentes possibilités. En fait, la métagrammaire n'est rien d'autre qu'un modèle linguistique raisonnable, c'est-à-dire sans les multiples redondances d'une grammaire TAG. Malheureusement, la métagrammaire n'a été développée que dans le but de générer une grammaire TAG et peut difficilement être utilisée pour autre chose.

Un formalisme comme celui de GUST permet d'écrire à la fois une grammaire complètement lexicalisée comme TAG et une grammaire modulaire comme la métagrammaire (voir Section 3.4). Et surtout, il permet de choisir comment répartir l'information linguistique entre les lexies si l'on décide de lexicaliser la grammaire. Par exemple, l'information concernant la place d'un clitique par rapport au verbe peut être attachée au verbe comme en TAG ou au clitique comme je le préconise dans [C10] ou [A3].

1.4 La grammaire syntagmatique dirigée par les têtes (HPSG)

Plutôt qu'une source d'inspiration, HPSG (Pollard & Sag 1994, Sag & Wasow 1999) se situe en point de mire de GUST, car c'est à mon avis, à l'heure actuelle, la plus développée des théories linguistiques complètement formalisées.

HPSG est également une grammaire lexicalisée dans le sens où une grande partie de l'information est attachée aux unités lexicales. Néanmoins, HPSG possède une architecture plus riche que TAG et les descriptions ne sont pas directement attachées aux unités lexicales. La grammaire repose entièrement sur une ontologie sous forme d'un treillis d'héritage (multiple) dont les unités lexicales sont les feuilles : les descriptions (appelées contraintes) sont attachées aux nœuds du treillis, et chaque nœud hérite de l'ensemble des descriptions de ses ancêtres, qui, unifiées entre elles, donnent la description effective du nœud.

HPSG n'est pas une grammaire complètement lexicalisée, car les descriptions des mots ne se combinent pas directement entre elles. HPSG possède un certain nombre de schémas de combinaison qui sont exprimés dans le même formalisme que les descriptions des mots. Ces schémas décrivent la combinaison d'une tête et d'un argument (*head-complement-phrase*, *head-subject-phrase*, …), d'une tête et d'un modifieur (*head-modifier-phrase*), d'une tête et d'un élément extrait (*head-filler-phrase*) ou de deux conjoints (*coordinate-phrase*). La combinaison de deux mots est en fait la combinaison de leurs descriptions avec un schéma de construction.

La répartition de l'information linguistique entre descriptions de mots et schémas est formellement libre. Ginzburg & Sag 2000, inspirés par les grammaires de constructions (Goldberg 1995), préconisent d'introduire des schémas différents pour des "constructions"

différentes, notamment pour les différents type d'extraction (*wh-rel-cl*, *wh-int-cl*, …). Dans [A7], je prends le contre-pied et montre qu'il est possible, et plus économique, de traiter l'extraction en évitant complètement de mettre dans les schémas de combinaison de l'information linguistique. Les schémas se réduisent alors à des opérations purement formelles comme le sont la substitution et l'adjonction en TAG et la grammaire HPSG devient proche d'une métagrammaire TAG[9] ou d'une grammaire de dépendance du type GUST.

Bien que HPSG ne cherche pas précisément à isoler des niveaux de représentation séparés, on peut admettre que HPSG considère à peu près les quatre mêmes niveaux de représentation que GUST : un niveau sémantique (traits SEM ou SYNSEM/CONT)[10], où sont indiqués entre autres les prédicats sémantiques et les relations prédicat-argument, un niveau syntaxique (traits SYN ou SYNSEM/CAT) où sont indiqués les relations de dépendance (sous-catégorisation, modification, …), un niveau topologique où sont indiqués l'ordre des mots et la structure topologique (trait DOM) et un niveau phonologique (trait PHON). Le niveau topologique a été introduit pour traiter les questions d'ordre des mots en allemand (Reape 1994, Kathol 1995 ; voir aussi Donohue & Sag 1999 pour le warlpiri). Mais, en général, la structure topologique n'est pas considérée et la structure de constituants est confondue avec la structure syntaxique comme dans la tradition chomskyenne.

Malgré ses qualités évidentes, HPSG pose quelques problèmes. Nous allons souligner ici deux points principaux sur lesquels GUST essaye d'apporter une réponse.

Premièrement, la distinction entre les différents niveaux structurels n'est pas toujours claire. Comme je l'ai mentionné ci-dessus, le niveau morphologique/topologique n'est en général pas considéré. HPSG met l'accent sur l'interface entre les niveaux de représentation, plutôt que sur les représentations elles-mêmes. Un même formalisme — les structures de traits typées — est utilisé pour décrire les différentes représentations et la structure mathématique sous-jacente à chaque type de représentation (graphe, arbre, suite, etc.) n'est pas mise en évidence. Le formalisme, équivalent aux machines de Turing, permet bien sûr d'encoder n'importe quel type de structure, mais en traduisant en HPSG une structure quelconque, comme par exemple un arbre à bulles, on perd en général l'accès immédiat aux propriétés géométriques de cette structure et aux généralisations qui en résultent.

Deuxièmement, la grammaire inclut des choix procéduraux. En fait, la structure syntagmatique d'une grammaire HPSG n'a plus de motivation théorique réelle. Dans les

---

[9] La génération d'une grammaire TAG à partir d'une grammaire HPSG a d'ailleurs été proposée par Kasper *et al.* 1995.

[10] Dans la volonté de coller davantage à la théorie saussurienne du signe linguistique à deux faces, Pollard & Sag 1994 considère que les mots et les syntagmes ont, en tant que signes, seulement deux traits principaux sous lesquels est enchâssé l'ensemble de la description : un trait PHON correspondant au signifiant phonologique et un trait SYNSEM correspondant au signifié et incluant les caractéristiques sémantiques et syntaxiques. Pollard & Sag 1987 et Sag & Wasow 1999 considèrent eux des traits SEM et SYN primitifs. Formellement, les deux approches sont équivalentes.

grammaires chomskyennes, la structure syntagmatique est utilisée pour encoder les relations de dépendance syntaxiques et les relations d'ordre entre les mots. Dans une grammaire HPSG, les relations de dépendance syntaxiques sont encodées dans les traits de valence des entrées lexicales (les traits COMP, SPEC et aussi MOD pour la relation entre un modifieur et son gouverneur syntaxique). Comme il est montré en [A7], la structure syntagmatique construite par une grammaire HPSG résulte uniquement de la procédure suivant laquelle sont combinés les mots : la procédure usuelle est une procédure de bas en haut (*bottom-up*), c'est-à-dire qu'un mot ne peut se combiner avec son gouverneur qu'après s'être combiné avec ses propres dépendants. Pour une phrase avec une topicalisation comme (1a), la procédure usuelle d'HPSG donne la structure de constituants (simplifiée) en (1b). On construit d'abord *Peter hates*, qui est reconnu comme un constituant non saturé, puisque *hates* requiert un objet direct. Cette requête est placée dans un trait spécifique, appelé SLASH, qui remonte lorsque le groupe *Peter hates* est combiné avec *you said*, puis *I know* ; enfin, le tout est combiné avec l'élément topicalisé *Sandy*, qui vient satisfaire la requête de *hates* placée dans le trait SLASH.

(1) a. *Sandy I know you said Peter hates*

b. [Sandy (I know [you said (Peter hates)])]

c. [([(*Sandy*) *I know*] *you said*) *Peter hates*]

On peut très bien construire des grammaires HPSG fortement équivalentes aux grammaires HPSG traditionnelles (c'est-à-dire construisant les mêmes représentations sémantiques, syntaxiques, topologiques et phonologiques) pour lesquelles la combinaison des mots suit une procédure différente et la structure syntagmatique n'est plus la même. Dans la dernière section de [A7], je propose une grammaire HPSG qui donne la structure de constituants (1c) et permet une analyse incrémentale de la phrase : l'élément topicalisé *Sandy* est combiné dès le départ avec *I know* et est placé dans un trait spécifique, appelé VISITOR, qui est le pendant du trait SLASH pour cette procédure. Le groupe *Sandy I know* est ensuite combiné avec *you said*, puis *Peter hates*, la requête de *hates* étant directement satisfaite par le contenu du trait VISITOR qui a été propagé. Mon propos n'est pas de dire que (1c) est une meilleur décomposition que (1b) (bien que (1c) permette une analyse plus incrémentale), mais que les deux analyses sont possibles et que toute analyse HPSG est liée à un parenthésage particulier.

A la différence d'HPSG, GUST ne spécifie pas l'"ordre" dans lequel les mots doivent être combinés et un mot peut se combiner avec son gouverneur à n'importe quel moment, qu'il se soit ou non déjà combiné avec certains ou la totalité de ses dépendants. Autrement dit, les deux procédures précédentes sont compatibles avec la même grammaire GUST, celle-ci ne spécifiant aucune procédure par elle-même (et aucun parenthésage particulier).

Dans la Section 2, nous aurons l'occasion de souligner un autre problème posé par HPSG concernant le traitement des signes grammaticaux.

1.5 La grammaire lexicale fonctionnelle (LFG)

Le dernier modèle par rapport auquel nous situerons GUST est LFG (Kaplan & Bresnan 1982, Bresnan 1998). A la différence d'HPSG, LFG possède deux niveaux de représentation bien

séparés, la f-structure, qui est un compromis entre un arbre de dépendance syntaxique et un graphe sémantique, et la c-structure, qui est une structure syntagmatique spécifiant l'ordre des mots. Dans [C19], nous proposons, avec Lionel Clément et Kim Gerdes, un fragment de grammaire LFG pour l'allemand. Nous montrons qu'il est possible d'obtenir une grammaire de l'allemand beaucoup plus élégante que les précédentes grammaires LFG de l'allemand (Zaenen & Kaplan 1995) en modifiant quelque peu les structures LFG : d'une part, en indiquant clairement les dépendances syntaxiques dans la f-structure (ce qui n'était pas fait pour l'instant) et d'autre part, en dégageant complètement la c-structure des questions de dépendance (en renonçant en particulier à des notations basées sur la Syntaxe X-barre) et en donnant plutôt à la c-structure le rôle d'une structure topologique (comme la structure morphologique de GUST). Le principal problème de LFG vient du fait que la f-structure contient deux structures en une (nos représentations sémantique et syntaxique) et qu'elle ne permet pas de distinguer simplement les unités sémantiques des unités syntaxiques, alors que les deux ne se correspondent pas une à une. Ce point (fondamental) mis à part, on peut adopter, à des modifications mineures près telles que l'introduction de la notion de champ, le formalisme de LFG pour modéliser l'interface syntaxe-morphologie, comme nous l'avons fait dans [C19].

Nous concluons la Section 1 ici. Nous n'avons évidemment pas fait un tour complet des modèles linguistiques existants. Néanmoins, TAG, HPSG et LFG constituent à eux trois un inventaire raisonnable des types de formalismes actuellement utilisés en modélisation des langues. Il s'agit également de trois des rares modèles pour lesquels des grammaires à large couverture ont été développées, ainsi que des analyseurs raisonnablement performants.

Je vais poursuivre ce travail d'analyse des modèles linguistiques en m'attaquant maintenant à la question du traitement des signes linguistiques lexicaux et grammaticaux. La théorie des signes développée dans la Section 2 constitue la base de l'architecture de GUST, laquelle sera exposée de façon formelle dans la Section 3.

2 A propos des signes linguistiques

Aborder la question des signes, c'est aborder une question essentielle et incontournable de la modélisation mathématique de la langue[11], à savoir la question des particules élémentaires de la modélisation, des briques avec lesquelles est construit le modèle d'une langue.

La plupart des modèles linguistiques actuels se sont développés à partir de modèles syntaxiques, notamment la Syntaxe X-barre (Jackendoff 1977). Avec la prise de conscience de l'importance des variations lexicales dans les années 1970, les modèles qui ont émergé à la fin des années 1970 comme LFG, TAG et G/HPSG ont mis en avant la nécessité d'un lexique informationnellement conséquent et structuré en conséquence. L'architecture du lexique de

[11] C'est aussi aborder un domaine "sacré" depuis les travaux de Saussure et ses nombreuses exégèses. Je m'excuse d'avance de la grande naïveté avec laquelle je risque d'aborder un domaine qui n'est pas le mien.

ces modèles est maintenant la suivante : une première composante du lexique donne la description des lexèmes (= signes lexicaux) ; une deuxième composante du lexique spécifie un certain nombre d'opérations, appelées *relations lexicales*, qui permettent de dériver de chaque lexème les mots-formes qui lui corresponde. Cette architecture du lexique est en quelque sorte le troisième stade de la formalisation des signes grammaticaux dans les modèles linguistiques. Prenons l'exemple bien connu du passif. Dans Chomsky 1957 (et les modèles chomskyens qui ont suivi), le passif est traité comme une transformation entre deux structures syntaxiques. Avec le développement de la composante lexicale des modèles linguistiques, l'idée émerge naturellement de ramener le traitement du passif à une relation à l'intérieur du lexique (Wasow 1977, Bresnan 1982). Il ne s'agit pas encore d'une relation entre un lexème et un mot-forme, mais plutôt d'une relation entre deux mots-formes, la forme active et la forme passive d'un verbe.

Dans une modèle où la relation lexème - mot-forme est traitée à l'intérieur du lexique, les signes grammaticaux ou *grammes*[12] n'ont pas réellement de statut. On pourrait dire qu'ils correspondent aux opérations qui permettent de dériver les mots-formes des lexèmes, mais la question n'est pas abordée. Par exemple, en HPSG, mots-formes et lexèmes reçoivent explicitement le statut de signes linguistiques[13], mais la question de savoir si les relations lexicales sont aussi des signes n'est jamais posée[14]. De plus, le traitement des relations lexicales est généralement grossier : par exemple, dans Sag & Wasow 1999 : 189, le passage du lexème anglais EAT au mot-forme *eats* (la forme du présent de l'indicatif de la 3ème personne du singulier) est effectué par une seule opération, sans que soient distingués les grammes de voix, mode, temps et accord en personne-nombre (ce qui au passage est une importante perte de généralité, ce à quoi HPSG est généralement fort sensible).

Les grammes ne sont pas les seuls parents pauvres des modèles linguistiques. C'est le cas également des locutions et des collocations, qui font généralement l'objet d'un traitement séparé et complètement ad hoc. Par exemple, Sag & Wasow 1999 : 269 traite la locution KICK THE BUCKET 'mourir' comme une seule entité sans aucune structure interne et sans

---

[12] Nous conservons l'usage de la TST qui est d'appeler *grammèmes* les significations grammaticales, c'est-à-dire les signifiés des signes flexionnels. Nous appelons par conséquent ces derniers des *grammes*.

[13] Pour sa part, Mel'čuk 1993 ne considère pas exactement les lexèmes comme des signes, mais plutôt comme des ensembles de signes. Dans la phrase *le garçon regarde les chevaux*, les mots-formes sont **le**, **garçon**, **regarde**, **les** et **chevaux** et les lexèmes sont LE (deux occurrences correspondant à **le** et **les**), GARÇON, REGARDER et CHEVAL. Le mot-forme est un signe linguistique, mais un lexème est "un ensemble d'éléments spécifiques ayant un "noyau" commun sur le plan sémantique" (*ibid* : 99). Par exemple, REGARDER est défini comme l'ensemble des mots-formes {**regarde**, **regardes**, **regarde**, **regardons**, **regardez**, **regardent** ; **regarderai**, **regarderas**, … ; **regardé**, …, **regardées** ; **regardant**, **regarder**}.

[14] A vrai dire, la nature formelle des relations lexicales interdit de se poser la question : un signe est une association entre un signifié et un signifiant, c'est-à-dire entre deux niveaux de représentation différents, alors qu'une relation lexicale est une relation entre deux objets de mêmes niveaux, deux mots-formes ou un lexème et un mot-forme.

que ne soit saisi le fait que, au moins pour une acception du terme *mot*, le *kick* de *kick the bucket* et le *kick* de *kick the ball* ‘frapper la balle’ sont un seul et même mot.

Nous allons voir que le fait que grammes et locutions soient tous deux mal traités n’est pas un hasard. Avant de présenter notre traitement des grammes, des locutions et des signes en général (Section 2.3), nous allons rappeler brièvement le traitement du signe chez Saussure (Section 2.1) et nous attarder sur le traitement des signes proposé par Igor Mel’čuk (Section 2.2), qui préfigure celui de GUST. La Section 2.4, en guise de conclusion, statuera sur la question de la double articulation de la langue.

2.1 Le signe saussurien et la double articulation

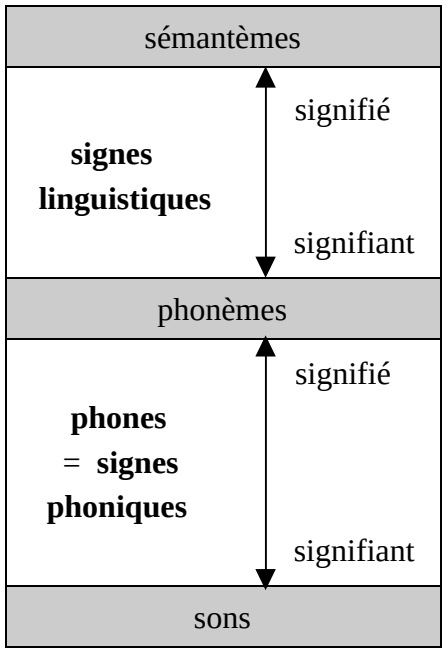
Comme chacun sait, les signes linguistiques sont des objets à deux faces — le signifié et le signifiant. Rappelons ce que dit Saussure du signifiant d’un signe linguistique (1916 : 98) : “Le signe linguistique unit non une chose et un nom, mais un concept et une image acoustique. Cette dernière n’est pas le son matériel, chose purement physique, mais l’empreinte psychique de ce son, la représentation que nous en donne le témoignage de nos sens ; elle est sensorielle, et s’il nous arrive de l’appeler «matérielle», c’est seulement dans ce sens et par opposition à l’autre terme de l’association, le concept, généralement plus abstrait.” Le signifiant du signe linguistique est ainsi décomposé en phonèmes : “le phonème est la somme des impressions acoustiques et des mouvements articulatoires, de l’unité étendue et de l’unité parlée, l’une conditionnant l’autre : ainsi c’est déjà une unité complexe qui a un pied dans chaque chaîne [la chaîne des mouvements de phonation vs. la chaîne parlée]” (*ibid.* : 65).

L’idée que le signifiant du signe linguistique n’est pas le son lui-même mais une combinaison d’unités plus abstraites a été reprise par Martinet (1960/1980 : 16) sous l’appellation de *double articulation* : “Comme tout signe le monème [l’unité de première articulation] est une unité à deux faces, une face signifiée, son sens ou sa valeur, et une face signifiante qui la manifeste sous forme phonique et qui est composée d’unités de deuxième articulation. Ces dernières sont nommées des phonèmes.”

Bien que le terme de *double articulation* suggère un parallèle entre les unités de première articulation — les signes — et les unités de deuxième articulation, Martinet, comme Saussure, n’envisage pas l’unité de deuxième articulation comme une unité à deux faces. Pourtant les phonèmes sont associés à des sons réels, comme les sémantèmes (les signifiés linguistiques) sont associés à des chaînes de phonèmes. Nous envisageons donc une unité à deux faces, que nous appellerons le *phone*, dont la face abstraite est le phonème et la face concrète le son[15]. A

[15] On utilise usuellement le terme *phone* pour désigner “les sons du langage, c’est-à-dire chacune des réalisations concrètes d’un phonème, variables suivant le contexte phonique, le locuteur, les conditions générales d’émission” (Dubois *et al.* 1994). De ce fait, le phone-son n’existe pas sans sa contrepartie, le phonème, et implique l’unité à deux faces (son,phonème). Je rappellerais à ce propos ce que dit Saussure du signe (p. 99) : “Nous appelons *signe* la combinaison du concept et de l’image acoustique : mais dans l’usage courant ce terme désigne généralement l’image acoustique seule, par exemple un mot (*arbor*, etc.). On oublie que si *arbor* est

partir de maintenant, nous utiliserons le terme *signe* pour désigner toute unité à deux faces et nous nous autoriserons à dire que le phonème est le signifié[16] du phone et que le phone est un *signe phonique*. De même, on peut considérer que dans la langue écrite interviennent des *signes graphiques* dont le signifiant est une certaine configuration de traits et le signifié une lettre ou un autre des symboles entrant dans la composition du signifiant d'un mot écrit. Nous réserverons le terme de *signe linguistique* ou de *signe saussurien* au signe dont le signifiant est une chaîne de phonèmes et le signifié un sémantème.



Dans la suite, nous ne reviendrons pas sur la question des phones et nous nous concentrons sur les différents types de signes que l'on peut considérer entre le niveau phonologique (le niveau des phonèmes) et le niveau sémantique (le niveau des sémantèmes). Le terme *signe* sera dorénavant utilisé dans l'acception large proposée ici.

2.2 Les signes linguistiques dans la TST

Igor Mel'čuk [dorénavant IM] consacre l'ensemble du volume 1 de son *Cours de Morphologie Générale* (Mel'čuk 1993) [dorénavant CMG] à ce que sont les objets élémentaires de la morphologie. IM prolonge la notion saussurienne de signe linguistique dans deux directions. D'une part, IM considère trois composantes pour un signe : le signifié,

---

appelé un signe, ce n'est qu'en tant qu'il porte le concept «arbre», de telle sorte que l'idée de la partie sensorielle implique celle du total."

[16] Parler du signifié du phone n'implique pas que le phone a un sens. Suivant Mel'čuk 1993, nous distinguons clairement *sens* et *signification*. Les phones ont une signification mais pas de sens. Ce ne sont pas des signes intentionnels, c'est-à-dire des signes choisis intentionnellement pour exprimer un sens. Ils ne sont choisis qu'indirectement, parce que leur signification entre dans la composition du signifiant d'un signe linguistique.

le signifiant et le syntactique, qui contient la combinatoire du signe et sur lequel nous reviendrons dans la Section 2.4. D'autre part, IM rompt avec l'idée que tous les signes linguistiques sont au même étage[17] entre un même plan du signifiant et un même plan du signifié. Nous allons étudier ce point maintenant.

Voici la citation la plus explicite d'IM (CMG : 115) : "De façon **générale,** est signifié linguistique tout élément linguistique qui correspond à un autre élément linguistique plus près de la surface phonétique. Symétriquement, est signifiant linguistique tout élément linguistique qui correspond à un autre élément linguistique plus près du sens. Alors, ce qui est un signifiant du niveau *n* (de la représentation de l'énoncé) pour un signifié du niveau *n-1*, peut être, à son tour, un signifié pour un signifiant du niveau *n+1*. Par exemple, une relation syntaxique de surface est un signifié de la construction qui l'exprime et, en même temps, un signifiant pour une ou plusieurs relations syntaxiques profondes."

Si IM évoque le fait que le signifié d'un signe puisse être à son tour le signifiant d'un signe d'un étage plus profond, il n'en donne qu'un exemple dans le CMG (p. 114) : "En dehors de la morphologie, un signifiant peut être, par exemple, une paire ordonnée d'ensembles de traits syntaxiques et de valeurs de variables morphologiques — c'est-à-dire une construction syntaxique de surface. Ainsi la construction $N_{\textbf{(genre)nombre}}$ + $Adj_{\textbf{genre,nombre}}$ est le signifiant de la relation syntaxique «modificative» (= «épithète») en français (*journal intéressant, revue intéressante, homme idéal, hommes idéaux*). Ce qu'on voit ici est une combinaison particulière de deux classes de mots (en l'occurrence, de deux parties du discours : un nom avec un adjectif) et manifestant des accords particuliers (en genre et en nombre). C'est cette combinaison de ces classes de mots dans cet ordre et avec ces accords qui constitue un signifiant. Cependant, dans le CMG nous ne considérons pas ce type de signifiant." IM justifie le fait de ne pas donner d'autres exemples par le fait que l'ouvrage est consacré à la morphologie et donc aux signes dont les signifiants composent le signifiant des mots-formes, mais l'idée de voir le signifié d'un signe comme le signifiant d'un signe plus profond n'est pas non plus explicitement exploitée dans les autres présentations qu'il a fait de la TST.

Par contre, IM donne des exemples de signes ayant des signifiés de niveaux différents (CMG : 112) : "Le signifié et le sens ne sont pas du tout la même chose. Le signifié d'une unité linguistique peut être, entre autres, une dépendance syntaxique, ou toute une classe de dépendances syntaxiques, ou bien un changement de la combinatoire d'une autre unité. Par exemple le signifié du suffixe de l'infinitif français (**-er**, **-ir**, **-oir**, **-re**) est la classe des constructions syntaxiques où l'infinitif est censé apparaître ; le signifié du suffixe d'un nom déverbatif (tel que **-ing** en anglais) est une instruction : remplacer (dans le syntactique du radical) la marque «verbe» par la marque «nom», ces marques (= parties du discours) spécifiant les classes de constructions syntaxiques qui acceptent un verbe ou un nom. Il n'en est pas moins vrai que la plupart des signifiés d'une langue sont des sens." On voit dans la

[17] Dans la suite, nous distinguerons clairement les *niveaux* et les *étages*, un étage correspondant à l'interface entre deux niveaux adjacents.

dernière phrase de cette citation, l'hésitation d'IM à rompre avec l'héritage saussurien et à envisager de façon systématique les signes comme des éléments d'interface entre des niveaux de représentation adjacents (comme proposé dans la citation ci-dessus extraite du CMG : 115) et à généraliser l'idée de signes dont le signifié ne serait pas un sens, ainsi que de signes dont le signifiant ne serait pas une chaîne phonologique ou une opération de niveau phonologique.

2.3 Les signes linguistiques dans GUST

Toutes les théories linguistiques s'accordent maintenant sur le fait qu'une langue est une interface entre des sens et des textes. J'en prendrais pour preuve cette citation de Brody 1997 qui constitue le premier paragraphe de son ouvrage consacré au Programme Minimaliste : "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." De là, on tire souvent, comme en HPSG, que les signes linguistiques sont tous au même étage, entre deux mêmes plans, le plan du signifié ou *niveau sémantique* et le plan du signifiant ou *niveau phonologique*.

Pour introduire la théorie des signes retenue pour GUST, je vais revenir sur l'exemple du passif (en français). Du point de vue du sens, le passif modifie la diathèse d'un verbe et donc la saillance de ses actants par rapport au fait qu'il exprime. Du point de vue de la forme, le passif s'exprime par le verbe **ÊTRE** suivi du **participe passé** du verbe avec promotion de l'objet initial en sujet et rétrogradation du sujet initial en complément d'agent. On voit bien qu'on ne souhaite pas parler de l'expression du passif en termes phonologiques. Du coup, on lui nie généralement le statut de signe (il est révélateur de voir à quel point le passif a pu recevoir de formalisations différentes, des transformations aux règles lexicales de LFG et G/HPSG). Or on peut très bien considérer le passif comme un *signe* que nous appellerons *profond* dont le signifiant n'appartiendrait pas au niveau phonologique, mais à un niveau plus profond, que nous appellerons le *niveau syntaxique*. (Les signes profonds sont soulignés. Les signes lexicaux sont en majuscule et les signes grammaticaux en minuscule.) De même, les éléments qui composent le signifiant du passif, **ÊTRE** et la flexion **participe passé**, peuvent être vus comme des *signes* que nous appellerons *intermédiaires*[18]. (Les signes intermédiaires sont écrits en gras.) Ces signes n'ont toujours pas leur signifiant au niveau phonologique, mais à un niveau intermédiaire entre le niveau syntaxique et le niveau phonologique, que nous appellerons le *niveau morphologique*. Le signe intermédiaire **ÊTRE** donne lieu, selon les signes intermédiaires grammaticaux avec lesquels ils se combinent, une série de *signes de surface* qui composent les formes de ce verbe : *SUIS, ES, EST, SOMMES*, etc. (Les signes de surface sont écrits en italique.) Alors que le signe intermédiaire **ÊTRE** désigne la lexie, le signe de surface *ÊTRE* désigne la forme infinitive de cette lexie. Ce dernier signe a pour signifié le signifiant de **ÊTRE** ⊕ **infinitif** et pour signifié la chaîne phonologique /ɛtʀə/. On

[18] En toute rigueur, ce ne sont pas des signes intermédiaires qui composent le signifiant d'un signe profond, mais leurs signifiés.

peut également considérer que ce signe n'est pas élémentaire et est le composé *ÉT-* ⊕ *-RE* où *ÉT-* est un signe que l'on retrouve dans *ÉTAIS* ou *ÊTES*[19] et qui a pour signifié le signifiant de **ÊTRE** et pour signifiant /ɛt/$^{\varepsilon \Rightarrow e}$ et *-RE* est un signe que l'on retrouve dans VENDRE et qui a pour signifié le signifiant du gramme **infinitif** et pour signifiant /ʀə/. Par contre, le signe de surface *SOMMES*, dont le signifié est le signifiant du composé des trois signes **ÊTRE** ⊕ **ind-présent** ⊕ **1** ⊕ **pl**, peut difficilement être décomposé (un tel signe est appelé un *mégamorphe* par IM). La question de la décomposition des signes de surface est un problème complexe que nous ne considérerons pas dans la suite. Dans GUST, nous traitons pour l'instant tous les signes de surface comme des mégamorphes, c'est-à-dire que nous ne traitons pas réellement la morphologie.

<table>
<tr><td colspan="5">niveau sémantique</td></tr>
<tr><td>signe
profond</td><td>passif</td><td>passé
composé</td><td colspan="2">ÊTRE
(copule)</td></tr>
<tr><td colspan="5">niveau syntaxique</td></tr>
<tr><td>signe
intermédiaire</td><td colspan="4">ÊTRE</td></tr>
<tr><td colspan="5">niveau morphologique</td></tr>
<tr><td>signe
de surface</td><td>SUIS</td><td>ES</td><td>SOMMES</td><td>ÊTRE</td></tr>
<tr><td colspan="5">niveau phonologique</td></tr>
</table>

La distinction entre les grammes profonds et les grammes intermédiaires est importante pour les temps verbaux aussi. Les temps composés comme le passé composé n'apparaissent qu'à l'étage profond. A l'étage intermédiaire, le passé composé d'un verbe X s'exprime par la combinaison de l'auxiliaire **AVOIR** (ou **ETRE**) au **présent** et du verbe **X** à la forme **participe passé**. Le même gramme intermédiaire **présent** apparaît à la fois dans l'expression du temps présent et du temps passé composé. Sans considérer au moins deux étages de signes, il est difficile de saisir cette généralisation.

Donnons un autre exemple : on a envie de dire que, au moins d'un certain point de vue, le mot *taureau* de la locution *prendre le taureau par les cornes* est le même mot *taureau* qui désigne l'animal. Or le *taureau* de la locution est un mot qui n'a pas de sens propre. La solution est simple : la locution PRENDRE LE TAUREAU PAR LES CORNES est un signe profond (élémentaire) dont le signifiant est une combinaison (de signifiés) de signes intermédiaires

[19] L'alternance /e/⇒/ɛ/ est régulière dans la conjugaison des verbes français (*céder, cède* ; *révéler, révèle* ; etc.) et découle directement de la combinaison de certains radicaux avec certains affixes flexionnels.

PRENDRE ⊕ (**LE** ⊕ **masc** ⊕ **sg)** ⊕ (**TAUREAU** ⊕ **sg)** ⊕ **PAR** ⊕ (**LE** ⊕ **fém** ⊕ **pl)** ⊕ (**CORNE** ⊕ **pl)**[20]. Et il n'y a qu'un seul signe intermédiaire **TAUREAU** dont le signifié peut être le signifiant d'un signe profond ou une composante du signifiant d'une locution, et deux signes de surface *TAUREAU* et *TAUREAUX* qui correspondent à la combinaison de **TAUREAU** avec les grammes du **singulier** et du **pluriel**.

<table>
<tr><td colspan="4">niveau sémantique</td></tr>
<tr><td>signe
profond</td><td>TAUREAU1
'animal'</td><td>TAUREAU2
'signe du zodiaque'</td><td>PRENDRE LE TAUREAU PAR LES CORNES</td></tr>
<tr><td colspan="4">niveau syntaxique</td></tr>
<tr><td>signe
intermédiaire</td><td colspan="3">TAUREAU</td></tr>
<tr><td colspan="4">niveau morphologique</td></tr>
<tr><td>signe
de surface</td><td colspan="2">TAUREAU</td><td>TAUREAUX</td></tr>
<tr><td colspan="4">niveau phonologique</td></tr>
</table>

La théorie des signes que nous adoptons reflète l'architecture générale de GUST. Nous considérons quatre niveaux de représentation : le niveau sémantique, le niveau syntaxique, le niveau morphologique et le niveau phonologique. Entre ces quatre niveaux de représentation, nous avons donc trois modules ou, si l'on préfère, trois ensembles de signes :

1) le *module profond* ou *module sémantique*, qui est constitué des signes profonds et qui réalise l'*interface sémantique-syntaxe* ;
2) le *module intermédiaire* ou *module syntaxique*, qui est constitué des signes intermédiaires et qui réalise l'*interface syntaxe-morphologie* ;
3) le *module de surface* ou *module morphologique*, qui est constitué des signes de surface et qui réalise l'*interface morphologie-phonologie*.

Un quatrième module, non traité par GUST, est constitué des phones et réalise l'interface entre le niveau phonologique et le signal acoustique.

[20] Il est plus usuel d'écrire cette combinaison de signes **PRENDRE LE**$_{\textbf{masc,sg}}$ **TAUREAU**$_{\textbf{sg}}$ **PAR LE**$_{\textbf{fém,pl}}$ **CORNE**$_{\textbf{pl}}$, comme nous le verrons dans la Section 3.

niveau sémantique				
module profond	=	module sémantique	=	interface sémantique-syntaxe
niveau syntaxique				
module intermédiaire	=	module syntaxique	=	interface syntaxe-morphologie
niveau morphologique				
module de surface	=	module morphologique	=	interface morphologie-phonologie
niveau phonologique				

La correspondance entre une représentation sémantique et une représentation phonologique est assurée par la combinaison de différents signes de différents modules. Par exemple, la correspondance entre le sens ‘raser(moi,moi)’ et la forme de surface *je me suis rasé* (non considérés les grammes de mode et de temps) met en jeu la voix réfléchie dans le module profond, la lexie **SE** dans le module intermédiaire et le mot *ME* dans le module de surface. Le fait que le réfléchi doive être traité comme une voix en français est justifié, entre autres, par le fait que le réfléchi impose l’auxiliaire **ÊTRE** pour l’expression du passé composé (*Paul s’est rasé* vs. *Paul l’a rasé*), ainsi que d’autres contrastes comme dans les constructions causatives (*Marie fait se raser Paul* vs. **Marie fait le raser Paul*) ou l’impersonnel (*Il se rase plusieurs personnes* vs. **Il les rase plusieurs personnes*) (cf. Mel’čuk 2001 pour une argumentation détaillée). Le découpage en plusieurs étages permet une répartition plus harmonieuse de l’information et évite les redondances. En particulier le signe intermédiaire **SE** qui exprime la voix réfléchie se retrouve aussi comme expression du réciproque (*Pierre et Jean se battent*) ou dans le signifiant de locutions comme S’ENFUIR. Le signe intermédiaire **SE** se combine avec des grammes d’accord de nombre-personne et ceci quel que soit le signe profond auquel il correspond (*Nous nous rasons* ; *nous nous battons* ; *nous nous enfuyons*) et donne, en combinaison avec ces grammes d’accord, les signes de surface *ME*, *TE*, *SE*, *NOUS*, *VOUS*. Quant au signe de surface *ME*, il provient également de la combinaison de **MOI** avec le cas **accusatif** (*il me rase*), et appartient donc aussi au paradigme *JE*, *ME* et *MOI*. Pour une théorie qui ne considère que des signes au sens saussurien, le fait que *me* apparaissent dans deux paradigmes différents tout en ayant dans les deux cas des propriétés similaires est un

problème majeur, puisqu'on ne peut à la fois considérer deux mots-formes *me* distincts et saisir les similarités de comportements de ces deux mots-formes[21].

Revenons justement aux mots-formes. Dans notre architecture, un mot-forme comme ***regardons*** dans *Jour après jour, nous regardons l'eau du robinet couler* est vu comme la combinaison de douze signes élémentaires :

- trois signes de surface[22] : *REGARD-*, *-∅-* (*indicatif-présent*) et *-ONS* (*1ère personne-pluriel*) ayant pour signifiés respectifs /ʀ(ə)gard/, /∅/ et /õ/ ;
- cinq signes intermédiaires : la lexie **REGARDER**, les grammes intermédiaires de l'**indicatif** et du **présent** et les grammes d'accord de la **1ère personne** et du **pluriel** ;
- quatre signes profonds : le lexème profond REGARDER2 (numérotation du Petit Robert 1987) et les grammes profonds de l'indicatif, du présent, et de l'actif.

Le traitement des grammes d'accord mérite un commentaire. Nous avons considéré deux grammes intermédiaires d'accord pour le verbe. En effet, comme le spécifie la grammaire traditionnelle, le verbe s'accorde en personne et en nombre avec son sujet. Ces deux accords peuvent être traités par deux "règles" différentes et correspondent donc à deux grammes intermédiaires distincts : le verbe s'accorde en nombre avec un sujet nominal ou reçoit le gramme **singulier** si son sujet n'est pas nominal (*Que Pierre vienne **est** inquiétant* ; *Voler **est** un délit*) ; le verbe s'accorde en personne avec un sujet pronominal ou reçoit le gramme **3ème personne** si son sujet n'est pas pronominal (*Le livre **est** sur la table* ; *Voler **est** un délit*)[23]. Par contre, il y a bel et bien un unique signe de surface pour l'accord en nombre-personne : on ne peut en effet distinguer dans les signifiants des signes de nombre-personne (phonologiques /ə/, /ə/, /ə/, /õ/, /e/, /ə/ comme graphiques *–e, -es, -e, -ons, -ez, -ent*) la combinaison d'un signifiant pour le nombre et d'un signifiant pour la personne. A l'étage profond, la personne ne peut non plus être dissociée du nombre, car 'nous' n'est pas une pluralité de 'moi'. Notons pour finir que si nous n'avions pas considéré nos trois étages de signes, nous n'aurions jamais pu dissocier nombre et personne. Pire encore, si l'on suit la logique d'un unique étage de signe linguistique, on doit en arriver aux conclusions de Martinet (1980 : 105) : "Les signifiants **discontinus** tels que /nu…õ/ dans /nukurõ/ *nous courrons* résultent fréquemment de ce qu'on appelle l'accord : dans /lezanimopes/ *les animaux paissent* rapproché de /lanimalpè/ *l'animal paît*, le signe «pluriel» reçoit trois expression distinctes : /leza…/ au lieu de /la…/, /…mo…/ au lieu de /…mal…/ et /…pes/ au lieu de /…pè/ ; on dira, si l'on veut, que

---

[21] Ceci conduit certains à ne pas considérer le réfléchi comme une voix et à interpréter le *se* de *il se rase* comme un pronom réfléchi (comme le serait le pronom anglais *himself*), ce qui permet de postuler plus facilement un seul mot-forme *me*. Il va sans dire que cette solution ne va pas sans problème.

[22] Comme nous l'avons dit, dans l'architecture actuelle de GUST, nous ne traitons pas la morphologie et nous ne considérons qu'un seul signe de surface *REGARDONS* dont le signifiant est la suite des phonèmes /ʀ(ə)gardõ/.

[23] De nombreux modèles considèrent que les noms comme les pronoms possèdent un trait de personne. Ceci n'est pas justifié et on pourrait alors tout aussi bien postuler un trait de personne pour les verbes à l'**infinitif** comme *voler* dans *voler est un délit*.

le signifiant de «pluriel» est /-ez-/ accompagné d'une variante particulière des signifiants correspondants aux signifiés «animal» et «paître». Il n'y a là, bien entendu, qu'un seul monème de pluriel, celui dont le signifiant est simplement /-e-/ dans /lešamãž/ *les chats mangent*." Je suis effectivement d'accord sur le fait qu'il n'y a qu'un seul gramme profond de pluriel dans *les animaux paissent* (que l'on considère généralement porté par le nom), mais par contre, dans notre architecture, cet unique gramme profond donne trois grammes intermédiaires portés respectivement par le déterminant, le nom et le verbe, ayant chacun leur caractéristiques propre et faisant chacun l'objet d'une description séparée.

Notre système de signes permet d'étendre à tous les signes, lexicaux comme grammaticaux, l'idée qui sous-tend les grammaires lexicalisées comme TAG ou HPSG. Pour ces dernières, la description d'une phrase est obtenue pas la combinaison des descriptions des mots de la phrase. Pour GUST, la description d'une phrase est obtenue par la combinaison de signes lexicaux et grammaticaux de différents modules, c'est-à-dire par la combinaison de signes profonds, intermédiaires et de surface lexicaux et grammaticaux. Alors que, en TAG, HPSG ou LFG, les différentes formes d'un verbe sont obtenues au niveau du lexique, en GUST, pour obtenir une forme verbale passive au passé composé de l'indicatif, il suffit de *composer* les descriptions du lexème verbal, de l'indicatif, du passif et du passé composé. Et si l'on veut la description de la même forme avec un actant relativisé et un autre cliticisé (*[la personne] par qui le livre lui a été donné*), il suffit de composer la forme verbale avec un pronom relatif et un pronom clitique (*qui* et *lui* dans notre exemple). Une telle façon de faire paraît simple et naturelle et elle l'est. Pourtant, il n'existe pas à ma connaissance de modèle linguistique qui fonctionne comme cela. Les modèles qui ne considèrent qu'un seul type de signe sont confrontés à des choix insolubles pour traiter tous les mots non "standard" : faut-il traiter les clitiques comme des mots à part entières ou des affixes (voir Miller & Sag 1993 pour le traitement des clitiques comme affixes en HPSG) ? ; faut-il traiter les auxiliaires comme des lexèmes à part entière ou comme des éléments d'une forme verbale complexe (voir Bresnan 1982 pour un traitement séparé des auxiliaires en LFG, abandonné dans Bresnan 1998) ? ; comment traiter les éléments d'une locution ? ; etc. En GUST, les clitiques sont avant tout des signes ; l'opposition mot-affixe n'intervient que là où elle pertinente, c'est-à-dire au niveau morphologique, où est traité l'ordre des mots (alors qu'en HPSG ou en LFG, l'opposition mot-affixe va décider si les clitiques sont traités comme des signes à part entière ou sont introduits au niveau du lexique par des règles lexicales). En GUST, les auxiliaires ou les locutions ne créent pas non plus de choix capitaux : l'auxiliaire est traité à la fois comme élément d'une forme verbale complexe (par le module profond) et comme lexème à part entière (par le module intermédiaire) et la locution est traitée à la fois comme un tout (par le module profond) et comme une combinaison de mots (par le module intermédiaire).

2.4 La question du syntactique et l'articulation de la langue

Nous allons revenir sur la question du syntactique soulevée par Igor Mel'čuk dans le CMG. IM considère, à juste titre, que la combinatoire d'un mot-forme n'est pas contrôlée par la seule combinatoire de son signifié et de son signifiant, c'est-à-dire par des propriétés

sémantiques et phonologiques. Interviennent aussi dans la combinatoire des mots-formes des propriétés syntaxiques et morphologiques (ou topologiques). Ces propriétés, qu'on ne peut attribuer ni au signifié, ni au signifiant du signe linguistique, forment une composante spécifique du signe que IM appelle le *syntactique*.

IM dit du syntactique (CMG : 120) : "La présence des syntactiques constitue une propriété spécifique des langues naturelles […]. Cette propriété les oppose à tous les autres systèmes de communication, codes, langages formels, etc. En effet, les langages formels — du code de la route jusqu'aux langages de logique formelle, de programmation, etc. — n'ont pas de syntactiques. Bien sûr, ils possèdent une syntaxe, c'est-à-dire des règles spécifiant les expressions correctes (= bien formées), mais ces règles ne doivent pas mentionner que le sens ou la forme des symboles intéressés. Dans un langage formel, on n'a jamais affaire à des traits capricieux et «illogiques» de symboles individuels qui ne seraient pas conditionnés par le sens ou la forme de ces symboles et qui seraient exclusivement maintenus grâce à la tradition de l'usage. La cooccurrence des symboles dans un langage formel est toujours standard, dépendant soit du sens, soit de la forme. Par contre, dans une langue naturelle, une partie importante de la cooccurrence des unités n'est pas standard, en ce sens qu'elle ne dépend ni du sens ni de la forme. Ce sont les données sur cette cooccurrence non standard qui constituent le syntactique des signes linguistiques."

Les signes de GUST — les signes profonds, intermédiaires et de surface — n'ont pas de syntactique à part. Leur combinatoire est entièrement contrôlée par la combinatoire de leur signifié et de leur signifiant. Par contre si nous considérons un mot-forme, dont le signifié et le signifiant n'appartiennent pas à des niveaux adjacents, il devient nécessaire de considérer un syntactique pour prendre en compte la combinatoire du signe à des niveaux intermédiaires. Par exemple, la combinatoire du mot-forme *regardons* va être contrôlée par la combinatoire de chacun des signes qui entre dans sa décomposition (REGARDER2, indicatif, présent, actif, **REGARDER**, **présent**, **1ère personne-pluriel**, *REGARD-*, *-∅-*, *-ONS*) et en particulier par la combinatoire des signifiés et signifiants, des niveaux syntactiques et morphologiques, qui ne sont pris en compte ni dans la combinatoire du signifiant de *regardons* (qui concerne le seul niveau phonologique), ni dans la combinatoire de son signifié (qui concerne le seul niveau sémantique).

Plutôt que de dire comme IM que la nécessité d'introduire les syntactiques constitue une propriété spécifique des langues, on peut dire que la particularité des langues naturelles est avant tout d'être *articulées*, c'est-à-dire de faire intervenir des niveaux de représentation intermédiaires où est contrôlée la combinatoire des signes, ne permettant pas ainsi de contrôler la combinatoire des signes avec les seuls niveaux extrêmes, les niveaux sémantique et phonologique. Ainsi GUST repose-t-il sur une triple articulation du signe — signes profonds, signes intermédiaires, signes de surface —, qui, avec les phones, nous donnes une quadruple articulation de la langue.

Les questions qui ont été abordées dans cette section, sur la nature des signes linguistiques, sont évidemment au cœur de mes travaux, même si c'est dans ce document qu'elles sont

isolées pour la première fois. Dans la section suivante, consacrée à la présentation de détail de GUST et de son formalisme, nous retrouverons l'architecture défendue ici, avec trois modules — modules sémantique, syntaxique et morphologique — correspondant aux trois ensembles de signes introduits ici — signes profonds, intermédiaires et de surface. Par ailleurs, dans la Section 6, nous aurons l'occasion de discuter d'un cas particulièrement intéressant de signes profonds : les fonctions lexicales.

3 Grammaire d'Unification Sens-Texte

Nous n'allons pas faire ici une présentation complète de GUST. L'objectif de cette section est plutôt de montrer en quoi l'architecture générale de GUST est originale et se distingue des autres modèles de la langue. Pour une présentation plus complète, je renvoie à [D1]. On notera des différences de terminologie notables entre la présente présentation et [D1], dues en particulier à la meilleure séparation entre signes et niveaux de représentation (voir Section 3.1).

La Section 3.1 présente rapidement les niveaux de représentation considérés par GUST (et repris de la TST). La Section 3.2 propose une classification des modélisations formelles de la langue et explore le lien entre les différents types de formalisation afin de montrer comment rendre compatibles des approches a priori aussi divergentes que TAG, HPSG ou la TST. La section 3.3 est consacrée au formalisme de GUST et propose un fragment de modèle pour le français. La Section 3.4 aborde l'utilisation de GUST pour l'analyse et la génération et la question de la lexicalisation.

3.1 Les différents niveaux de représentation de GUST

Nous allons présenter les 4 niveaux de représentation de GUST (repris à la TST) : la représentation sémantique, la représentation syntaxique (de surface), la représentation morphologique et la représentation phonologique.

La représentation sémantique

L'objectif de la représentation sémantique TST est de représenter le *sens linguistique* d'un énoncé, c'est-à-dire l'organisation des signifiés des signes linguistiques apparaissant dans l'énoncé. Le signifié d'un signe linguistique possédant un sens est appelé un *sémantème*[24]. Les

[24] Nous considérons, à la suite de Mel'čuk, qu'un sémantème est a priori le signifié de plusieurs signes. Par exemple, le verbe VOYAGER en (i-a) et le nom VOYAGE en (i-b) possèdent des signifiés communs, de même que l'adverbe PENDANT en (i-a) et le verbe DURER en (i-b).

(i) a. *Pierre a voyagé pendant deux semaines.*
b. *Le voyage de Pierre a duré deux semaines.*

sémantèmes sont notés entre guillemets simples ; les sémantèmes lexicaux sont écrits en minuscules ordinaires (‘essayer’, ‘soupe’, etc.) et les sémantèmes grammaticaux en italiques (‘*présent*’, ‘*singulier*’, etc.)[25]. La Figure 1 donne une représentation sémantique de la phrase (2) simplifiée où apparaissent seulement les sémantèmes de la phrase et les relations prédicat-argument entre ces sémantèmes.

(2) Zoé essaye de manger la soupe

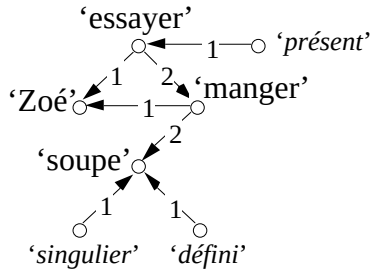


Figure 1 La représentation sémantique de (2)

Dans la Figure 1, les relations prédicat-argument ou dépendances sémantiques sont représentés par des arcs, mettant en évidence la géométrie de la représentation sémantique sous forme d’un graphe orienté. Les numéros sur les arcs indiquent la position de l’argument dans la liste argumentale du prédicat. On peut également donner cette représentation sémantique sous forme logique. Ceci nécessite de réifier les prédicats, c’est-à-dire de leur attribuer un nom qui sera utilisé lorsqu’ils sont arguments. (Seuls les sémantèmes lexicaux sont réifiés.) La représentation de la Figure 1 devient :

(3) e : ‘essayer’(*x*,*e′*)
e′ : ‘manger’(*x*,*y*)
x : ‘Zoé’
y : ‘soupe’
‘*présent*’(*e*)
‘*défini*’(*y*)
‘*singulier*’(*y*)

La différence de sens entre les phrases (i-a) et (i-b), s’il y en a une, ne vient pas des signifiés lexicaux, mais des signifiés grammaticaux — la détermination sur VOYAGE en (i-b) qui n’a pas de contrepartie en (i-a) — et de la structure communicative, le thème ayant tendance en français à s’assimiler, surtout dans des phrases hors contexte comme celles-ci, avec le sujet syntaxique.

[25] Dans tout notre exposé, nous resterons très grossier sur les questions de sémantique. Ici, comme dans la suite, nous ne distinguons pas les différentes acceptions d’un lexème comme nous devrions. De plus, nous ne nous intéressons pas à la décomposition sémantique des sémantèmes. Nous préférons user de sémantèmes grammaticaux comme ‘*présent*’, ‘*défini*’, … plutôt que tenter une description de ce que recouvrent exactement ces significations grammaticales, ce qui n’est pas notre objet ici.

Nous limitons notre présentation de la représentation sémantique au minimum nécessaire à la compréhension de ce qui suit. On pourra consulter [A2] ou [D1] pour une présentation plus détaillée considérant les autres structures de la représentation sémantique.

La représentation syntaxique

La représentation syntaxique de la phrase est un arbre de dépendance dont les nœuds correspondent aux mots-formes de la phrase et sont étiquetés par un lexème accompagné d'une suite de grammèmes. Cette représentation est similaire à la représentation syntaxique de surface de la TST, avec une différence concernant l'interprétation donnée aux termes et *lexème* et *grammème*. Lexèmes et grammèmes sont pour moi des unités monoplanes qui sont à l'interface des signes profonds et intermédiaires : un *lexème* est le signifié d'un signe lexical intermédiaire et un *grammème* le signifié d'un signe grammatical intermédiaire. Notre définition n'est pas incompatible avec la définition de Mel'čuk, qui définit un lexème comme un ensemble de mots-formes distingués uniquement par la flexion (voir Note 13). Plutôt que de voir le lexème comme un ensemble de signes, nous voyons le lexème comme le signifié commun à un ensemble de signes ayant le même signifié au niveau syntaxique. Dans les deux cas, le lexème est une abstraction qui sert à sélectionner un ensemble de signes linguistiques[26] et cet ensemble est le même dans les deux cas. Les lexèmes sont notés en majuscules ordinaires (PARLER, DE, LE, etc.) et les grammèmes en minuscules ordinaires (indicatif, présent, singulier, etc.).

Les dépendances syntaxiques sont étiquetées par des relations syntaxiques ou *syntaxèmes*. Les relations syntaxiques peuvent être considérées, de la même façon que les lexèmes et les grammèmes, comme des signifiés de signes intermédiaires. Par exemple, la fonction *sujet* du français est le signifié de la combinaison de quatre signes intermédiaires : un syntagme[27] (la règle d'ordre entre le verbe et son sujet), deux grammes d'accord (les accords du verbe avec son sujet en personne et en nombre) et un gramme de régime (le nominatif sur les pronoms).

---

[26] En effet, le lexème vu comme signifié d'un signe intermédiaire sélectionne également un ensemble de signes linguistiques : il s'agit de l'ensemble de tous les signes linguistiques qui ont ce lexème pour signifié. Par exemple, le lexème PARLER va se retrouver dans le signifié de niveau syntaxique de toutes les formes du lexème PARLER : *parle, parles, parlons, parlera, parleraient,* etc. Le statut formel que nous donnons au lexème (celui de signifié d'un signe intermédiaire) ne modifie pas le rôle qui est assigné à cette notion dans un modèle linguistique. Nous cherchons seulement à modéliser cette notion d'une façon à la fois plus élégante et plus proche de l'intuition que nous en avons. Personnellement, je ressens le lexème comme une unité de la description linguistique et pas comme un ensemble d'objets.

[27] Suivant Mel'čuk 1993, nous utilisons le terme *syntagme* dans deux acceptions. Dans la première, traditionnelle, le terme *syntagme* (= angl. *phrase*) désigne une combinaison linéaire d'au moins deux mots-formes qui sont syntaxiquement liés. Dans la deuxième, utilisée ici, le terme *syntagme* désigne "un signe élémentaire dont le signifié est la relation syntaxique exprimée, le signifiant est constitué de la combinaison de deux classes de mots, de l'ordre, des valeurs morphologiques et des prosodies correspondantes qui expriment cette relation" (*ibid.* : 129). Contrairement à Mel'čuk, je pense que le signe exprimant la relation d'ordre peut et doit être séparé des signes exprimant les accords et le régime (les grammes) et je réserve le terme *syntagme* à ce seul signe.

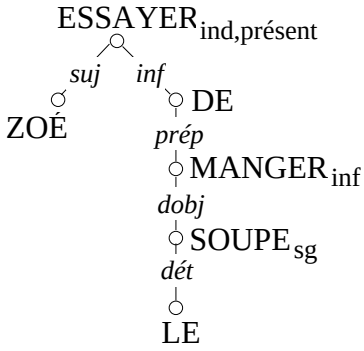


Figure 2 La représentation syntaxique de (2)

La représentation morphologique

La représentation morphologique de la phrase est la suite des représentations morphologiques des mots de la phrase. Comme beaucoup d'autres modèles utilisés en traitement automatique de la langue, GUST ne traite pas la morphologie et utilise un ensemble de règles qui associent directement les représentations morphologiques des mots à leur représentation phonologique ou graphique. En raison de ce traitement brutal du module morphologique, la nature de la représentation morphologique adoptée n'est pas très déterminante. Nous donnons en (4) la représentation morphologique profonde de (2) et en (5) sa représentation morphologique de surface. Nous nous permettons d'appeler *morphes* les signes de surface et *morphèmes* leurs signifiés. Le morphème est donc une unité monoplane, qui entre aussi dans la composition des signifiants des signes intermédiaires. La représentation morphologique de surface d'une phrase donne la chaîne des morphèmes (notés entre accolades), tandis que la représentation syntaxique profonde est une version linéarisée de la structure syntaxique, dans laquelle figurent encore les lexèmes et les grammèmes. La représentation morphologique de surface pose un problème : là où nous avons un seul morphème, comme {3SG} ou {IND.PRÉ}, nous avons deux grammèmes. Autrement dit, la partition de la représentation morphologique effectuée par les signes de surfaces (les morphes) est moins fine que la partition effectuée par les signes intermédiaires (les grammes). Pour cette raison, dans l'état actuel de GUST (où la morphologie n'est pas traitée), nous adoptons comme représentation morphologique la représentation morphologique profonde de la TST.

(4) *ZOÉ ESSAYER$_{ind,présent,3,sg}$ DE MANGER$_{inf}$ À LE$_{fém,sg}$ SOUPE$_{sg}$*

(5) {ZOÉ} {ESSAY-}.{IND.PRÉ}.{3SG} {DE} {MANG-}.{INF} {LE}.{FÉM}.{SG} {SOUPE}.{SG}

Il n'est pas nécessaire que nous présentions la représentation phonologique, dont le squelette est une suite de phonèmes.

Comme le terme *lexème*, le terme *morphème* est souvent utilisé pour désigner des signes ou des ensembles de signes. Nous proposons de systématiser l'usage des termes en *-ème* et d'utiliser ceux-ci uniquement pour désigner des unités monoplanes représentant des signifiés et/ou des signifiants. Plus précisément, nous appellerons *Xème* le signifié d'un *Xe* et *Xe* un signe dont le signifié est un *Xème*. Ainsi appellerons-nous *lexe* un signe intermédiaire lexical

(dont le signifié est un lexème) et *sémante* un signe profond (dont le signifié est un sémantème). Il est d'usage d'appeler *lexie* un sémante lexical dont le signifiant est un lexème ou une configuration de lexèmes. Par parallèlisme, on appellera *grammie* un sémante grammatical et on réservera le terme *gramme* à un signe grammatical intermédiaire.

Nous terminons cette section par un tableau récapitulatif des termes introduits jusqu'ici.

sémantème		
sémante (lexie, grammie)	=	signe profond
lexème, grammème, syntaxème		
lexe, gramme, syntagme	=	signe intermédiaire
morphème		
morphe	=	signe de surface
phonème		
phone	=	signe phonique
son		

3.2 Grammaires transductives, génératives et équatives

Avant de pouvoir exploiter dans un même formalisme des idées venant de formalismes aussi éloignés que les grammaires de correspondance de la TST, les grammaires génératives[28] comme TAG ou les grammaires basées sur les contraintes comme HPSG, un travail de comparaison et de classification est nécessaire. Nous allons présenter ici le travail développé dans [C12].

Nous partons de la TST. Les règles de la TST sont des règles qui mettent en correspondance deux fragments de structures appartenant à des niveaux de représentation adjacents. Par exemple, un fragment de graphe sémantique avec un fragment d'arbre syntaxique comme

[28] Par *grammaire générative*, j'entends une grammaire qui permet de générer un ensemble de structures, par exemple un ensemble de suites comme les grammaires de réécriture de Chomsky ou un ensemble d'arbres comme les TAG. Le terme *grammaire générative* est souvent employé dans une autre acception pour désigner les modèles de l'école chomskyenne.

dans la Figure 3. Ou un fragment d'arbre syntaxique avec un fragment de chaîne morphologique (c'est-à-dire un fragment d'ordre linéaire) comme dans la Figure 4.

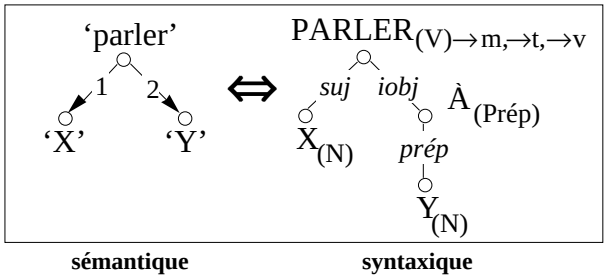


Figure 3 Une règle de correspondance TST (interface sémantique-syntaxe)[29]

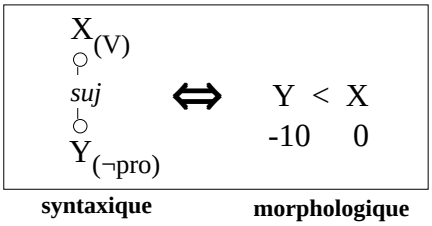


Figure 4 Une règle de correspondance TST (interface syntaxe-morphologie)[30]

Etant donnés deux ensembles *S* et *S'* de structures (graphes, arbres, suites, …), nous appellerons *grammaire de correspondance* ou *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, où une *règle de correspondance* est une règle qui met en correspondance un fragment d'une structure de *S* avec un fragment d'une structure de *S'*. 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.[31]

---

[29] La règle de la Figure 3 n'est pas tout à fait une règle TST standard. D'une part, il s'agit d'une règle mettant en correspondance directement une portion de graphe sémantique avec une portion d'arbre syntaxique de surface, sans passer par la structure syntaxique profonde. D'autre part, nous avons ajouté les traits →m, →t, →v qui doivent assurer que le verbe sera bien combiné avec des grammes de mode, temps, voix. Cette règle correspond au signe lexical profond PARLER de GUST.

[30] La règle de la Figure 4 n'est pas non plus une règle TST standard. Cette règle correspond au syntagme ou signe intermédiaire ***sujet_canonique***. D'autres syntagmes ont pour signifié la relation syntaxique *sujet* : ***sujet_inversé*** et ***sujet_clitique*** (cf. [A3]). La règle de la Figure 4 spécifie qu'un Y sujet de X peut se placer devant X à la distance –10, c'est-à-dire en laissant 9 positions entre lui et X pour d'autres dépendants de X. Plus exactement, ce n'est pas X et Y qui sont placés, mais les morphèmes {X} et {Y} correspondant aux lexèmes X et Y.

[31] La composée de deux grammaires transductives ne donne pas a priori une grammaire transductive. La difficulté vient du fait que, si G est une grammaire transductive entre *S* et *S'* et G' une grammaire transductive entre *S'* et *S"*, les fragments de structure de *S'* considérés par G ne sont pas forcément les mêmes que ceux considérés par G'. Ainsi l'interface sémantique-syntaxe de la TST considère comme fragments des portions

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* ([C12]). Par exemple, le module sémantique de la TST ne fait pas que mettre en correspondance des graphes sémantiques et des arbres de dépendance. La Figure 5 montre un graphe sémantique et un arbre syntaxique qui se correspondent, avec les partitions des deux structures qui en résulte. Par exemple, la règle qui associe 'essayer' à ESSAYER (c'est-à-dire le signe profond ESSAYER) va aussi introduire la préposition DE et le grammème infinitif.

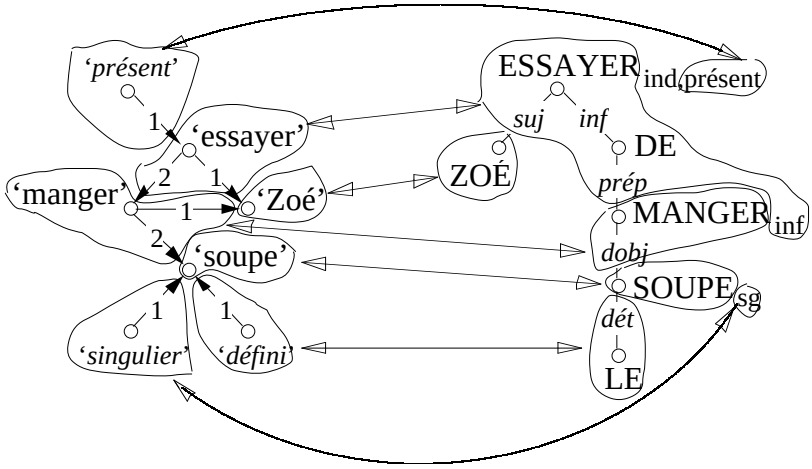


Figure 5 Correspondance entre une partition de graphe sémantique et une partition d'arbre syntaxique[32]

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'.[33] 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

---

importantes d'arbre syntaxique qui correspondent dans la structure sémantique à un seul nœud (par exemple pour une locution). Par contre, s'il existe une partition des structures de *S*' qui subsument à la fois les partitions introduites par G et G' (avec une condition de finitude sur les partitions), alors on peut construire, en composant les règles de G et G', une grammaire transductive G∘G' qui définit la correspondance entre *S* et *S''* obtenue par composition des correspondances définies par G et G'. Voir la Section 3.4 sur la lexicalisation.

[32] Un élément de la représentation syntaxique ne correspond pas à un élément visible de la représentation sémantique : il s'agit de l'indicatif sur le verbe ESSAYER qui résulte du fait que 'essayer' a été choisi pour exprimer la racine de l'arbre syntaxique. Ce choix n'est pas commandé par le graphe sémantique (les sémantèmes et les relations prédicat-argument entre eux), mais par la structure communicative, qui n'est pas représentée ici.

[33] Une simple correspondance entre *S* et *S'* est mathématiquement équivalente à un ensemble de couples (s,s').

obtenue par l'enchevêtrement des structures s et s' (l'enchevêtrement est dû au fait que, en un sens, les deux structures sont définies sur le même ensemble – l'ensemble des fragments mis en correspondance). Un tel triplet (s,s', $\varphi_{(s,s')}$) nous est donné par laFigure 5 où s est le graphe sémantique de gauche, s' l'arbre syntaxique de droite et $\varphi_{(s,s')}$ la correspondance entre les partitions de s et s'. On peut représenter la même information par la Figure 6.

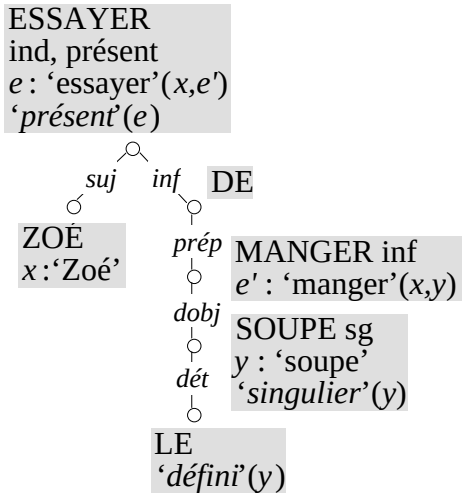


Figure 6 : Une structure produit équivalente à la Figure 5

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 (cf. Figure 7). C'est cette idée qui est à la base de la formalisation de GUST.

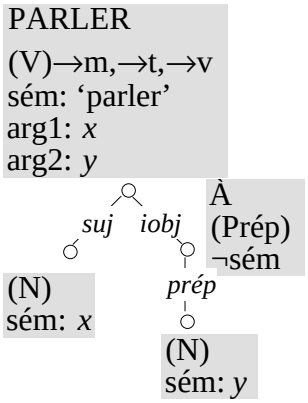


Figure 7 Un règle à la GUST (le signe profond PARLER)[34]

Avant de poursuivre la présentation de GUST, revenons aux grammaires génératives. Les grammaires génératives qui génèrent des structures produits, et plus généralement toutes les grammaires qui définissent des structures produits, peuvent être vues comme des grammaires

[34] Nous adoptons les notations de GUST sous forme de structure de traits : un sémantème est introduit par le trait sém et ses arguments par les traits arg1, arg2, etc.

transductives. C'est à mon avis le cas de tous les modèles utilisés en modélisation des langues. Par exemple, les modèles dont nous avons parlé jusqu'à présent définissent bien des structures produits : LFG, qui associe à une suite de mots une *c*-structure et une *f*-structure, TAG qui associe à une suite de mots un arbre syntagmatique (arbre dérivé) et un arbre de dépendance (arbre de dérivation) et HPSG, qui associe une chaîne phonologique à un structure de traits sémantico-syntaxique (les deux étant entremêlées du fait qu'il s'agit d'une structure produit). Ce n'est évidemment pas le cas de n'importe quelle grammaire générative ; il paraît en effet difficile d'interpréter une machine de Turing quelconque comme une grammaire transductive, car on voit mal quelle structure associer à une suite générée par la machine. Notons que certaines grammaires génératives peuvent s'interpréter comme des grammaires transductives, même si elles ne considèrent pas clairement des niveaux de représentations indépendants. Par exemple, dans le cas des grammaires de réécriture hors-contexte, ce n'est pas le résultat de la dérivation (qui n'est qu'une suite) qui est interprété comme une structure produit, mais le processus de dérivation lui-même (l'arbre de dérivation)[35]. Le fait que ces grammaires définissent une supercorrespondance n'est en fait qu'un effet de bord. Il me semble que le flou qui entoure l'une des notions les plus importantes de la grammaire générative, la *capacité générative forte*[36], vient du fait que les grammaires génératives ne sont pas clairement présentées comme mettant en correspondance des structures de différents niveaux, alors que c'est justement de ce point de vue qu'elle intéresse la modélisation des langues et qu'on veut les comparer. Je pense en effet que deux grammaires sont *fortement équivalentes* si elles sont équivalentes en tant que grammaires transductives, c'est-à-dire si elles définissent la même supercorrespondance. Et que les grammaires génératives n'ont d'intérêt en modélisation des langues que dans la mesure où elles définissent une supercorrespondance et simulent une grammaire transductive.

Maintenant qu'il est acquis que les seules grammaires génératives qui nous intéressent sont celles qui génèrent une supercorrespondance, revenons sur ce qui les distingue des grammaires transductives. Il s'agit bien évidemment de la procédure selon laquelle est définie la supercorrespondance à partir des règles. Dans tous les cas, la grammaire définit une supercorrespondance entre deux ensembles de structures. Dans la *présentation transductive* d'une grammaire, on se donne une structure de l'un des deux ensembles et l'on montre comment la grammaire permet de lui associer une structure correspondante. Dans la *présentation générative*, on ne se donne rien au départ et l'on génère simultanément deux structures en correspondance. Il y a une troisième possibilité logique que nous appelons la *présentation équative* : on se donne deux structures, une de chaque ensemble, et l'on utilise les règles de la grammaire pour vérifier si ces deux structures se correspondent. La grammaire est alors utilisée comme un filtre ou un ensemble de contraintes à satisfaire (d'où la terminologie usuelle de *grammaires basées sur les contraintes* ou *constraint-based*

[35] On peut néanmoins fortement simuler une grammaire de réécriture hors-contexte G par une grammaire de combinaison d'arbres à la manière de TAG qui génèrera les arbres de dérivation de G.

[36] Cf. Miller 1999 pour un travail récent sur le sujet.

grammars). Dans [C12], je montre comment un même ensemble de règle définit la même supercorrespondance suivant les différentes procédures présentées ci-dessus. Il s'agit d'un exemple d'une *grammaire de dépendance*, c'est-à-dire d'une grammaire définissant une supercorrespondance entre arbres de dépendance et suites ; à travers cet exemple est présentée la plus élémentaire des familles de grammaires de dépendance, les règles se limitant à mettre en correspondance une dépendance avec une spécification élémentaire d'ordre.

La catégorisation des présentations d'une grammaire ne se limite pas à l'opposition génératif, transductif, équatif. Si l'on prend l'exemple d'une grammaire de dépendance associant arbres de dépendance et suites, il y a évidemment deux présentations transductives selon qu'on parte des arbres (le sens de la synthèse) ou des suites (le sens de l'analyse). Qui plus est, dans chaque cas, on aura tendance à s'appuyer sur la structure de départ : une présentation dans le sens de la synthèse sera naturellement guidée par la structure d'arbre, tandis qu'une présentation dans le sens de l'analyse sera plutôt guidée par la structure d'ordre linéaire. De plus, la procédure associée à la grammaire doit assurer que la structure de départ est bien complètement traduite et qu'elle est donc bien complètement parcourue. Pour cette raison, les présentations transductives dans le sens de la synthèse assurent généralement un parcours de l'arbre de la racine vers les feuilles (parcours *top-down*), et les présentations transductives dans le sens de l'analyse assurent généralement un parcours de la suite ordonnée d'une extrémité à l'autre (analyse incrémentale). On retrouve les mêmes particularités dans les présentations génératives. En général, la procédure de génération (= la dérivation) s'appuie sur l'une des deux structures en correspondance. Par exemple, la présentation des grammaires de réécriture hors-contexte s'appuie très clairement sur la structure d'arbre (et pas sur la structure d'ordre de la suite générée) et plus précisément sur un parcours de haut en bas (*top-down*).

Ces questions sont prises en compte dans la définition de GUST. GUST possède une propriété importante que nous appellerons l'*associativité* : les règles peuvent se combiner dans n'importe quel ordre. Plus précisément, étant données trois règles (de correspondance) A, B et C, si A et B peuvent se combiner et que C peut se combiner avec le résultat A ⊕ B de la combinaison de A et B, alors C peut également se combiner avec A ou B et le résultat ne dépend pas de l'ordre de combinaison : (A ⊕ B) ⊕ C = A ⊕ (B ⊕ C). (Nous utilisons le même symbole ⊕ pour noter la combinaison entre les règles et la combinaison entre les signes, les deux notions étant confondues dans GUST.) Cette propriété assure que la grammaire est la moins procédurale possible et que toutes les procédures imaginables sont compatible avec la grammaire. On peut alors envisager des procédures qui s'appuient sur n'importe laquelle des deux structures mises en correspondance par la grammaire : la grammaire est donc complètement réversible, pouvant être utilisée *directement* dans le sens de la synthèse comme de l'analyse[37]. Pour une telle grammaire, une procédure de dérivation *globale*, ne spécifiant

[37] On arrive toujours à utiliser dans un sens une grammaire prévue pour l'autre sens. Par exemple, on peut utiliser dans le sens de l'analyse les grammaires de réécriture hors-contexte, qui sont définies plutôt dans le sens de la synthèse (la procédure de dérivation étant guidée par la structure d'arbre et non par l'ordre linéaire). Néanmoins

aucun ordre particulier dans la combinaison des règles et n'étant orientée ni dans le sens de la synthèse[38], ni dans le sens de l'analyse, est possible. C'est de cette façon que nous présenterons GUST.

3.3 Le formalisme de GUST

Dans la Section 2, nous avons présenté en détail la théorie des signes qui sous-tend l'architecture de GUST. En GUST, les seules règles sont des signes. Un signe est une unité à deux faces. Un signe peut être vu comme une règle qui met en correspondance deux fragments de structures à la façon de la TST ou, ce qui revient au même, comme un fragment de structure produit. Les signes de GUST peuvent être combinés les uns aux autres à la manière des structures élémentaires d'une TAG lexicalisée.

La première proposition de grammaire d'unification basée sur la TST est due à Boyer & Lapalme 1985. Malheureusement, le travail n'a pas eu de suite et le formalisme est resté à l'état d'ébauche. Le formalisme proposé ici doit beaucoup au travail de Nasr 1996, qui est la première grammaire de combinaison d'arbres basée sur la TST. Chez Nasr, comme dans mes premiers travaux ([A3]), les structures élémentaires correspondent à des fragments de représentation plutôt qu'à des signes. Et le formalisme ne permet pas de combiner des structures de différents étages.

GUST est une grammaire d'unification, ce qui signifie que les structures élémentaires associées aux signes se combinent par unification. L'unification est une opération très simple : deux structures se combinent par unification si elles peuvent être en partie superposées l'une à l'autre ; les deux fragments superposés sont identifiés et le tout donne une nouvelle structure. Cette idée de superposition est exprimée par Kay 1979, 1984, qui fut probablement le premier à utiliser l'unification en modélisation des langues. Pour notre part c'est vraiment dans cet esprit que nous utilisons l'unification, en collant ensemble des bouts de structure pour obtenir une structure complète. En particulier, GUST ne fait pas usage de la percolation[39] de traits comme certaines grammaires d'unification et tout particulièrement HPSG.

---

une telle procédure n'a rien de *direct* et diverses solutions peuvent être envisagées (comme par exemple de mettre la grammaire sous forme normale de Gaifman-Greibach pour obtenir un analyseur à pile).

[38] Igor Mel'čuk insiste beaucoup sur le fait qu'un modèle linguistique doit être pensé et écrit dans le sens de la synthèse, parce que l'activité naturelle d'un sujet parlant est justement de parler et que certaines caractéristiques de la langue, comme les fonctions lexicales, sont surtout pertinentes pour la synthèse. Je partage ce point de vue que le modèle doit d'abord être pensé dans le sens de la synthèse. Ce n'est nullement incompatible avec le fait que la grammaire soit réversible.

[39] L'idée de la percolation de traits est la suivante. Dans une grammaire d'unification, il est possible de sous-spécifier la valeur d'un trait dans une structure. Lorsque la structure est combinée avec une autre, ce trait reçoit une valeur s'il est instancié dans l'autre structure. Jusque-là rien de grave. Maintenant, il est également possible de faire partager une valeur à deux traits d'une structure : lorsque l'un des traits est instancié, l'autre l'est aussi, par la même valeur. Si l'on combine une chaîne de structures avec les mêmes deux traits *t* partagés et sous-spécifiés, il peut arriver la chose suivante : lorsqu'on instancie l'un des traits *t* à un bout de la chaîne, la valeur se

Le fait d'utiliser l'unification présente quelques contraintes. La TST fait un grand usage de conditions dans la formulation des règles. Une règle TST a la forme suivante : X ⇔ Y | C, signifiant que le fragment de structure X peut donner le fragment de structure Y (et inversement) si les conditions C sont vérifiées. Le langage des conditions n'est pas contraint, mais dans la pratique, il s'agit de formules logiques du 1er ordre portant sur le contexte de X et Y. Dans une grammaire d'unification comme GUST, les seules conditions possibles sur la combinatoire d'un signe (l'équivalent d'une règle TST) sont exprimées directement dans la structure élémentaire de ce signe dans le but de restreindre l'unification de ce signe avec d'autres : pour cela, on peut soit élargir le contexte du signe pour réduire ses possibilités de combinaison, soit ajouter des traits pour bloquer certaines unifications. N'importe quelle condition ne peut être exprimée de cette façon, d'autant qu'on souhaite que les traits qu'on introduit pour contrôler la combinatoire d'un signe soient linguistiquement motivés.

GUST ne fait plus à proprement parler de distinction formelle entre grammaire et lexique et tous les signes, qu'ils soient lexicaux ou grammaticaux, sont exprimés dans le même formalisme. Les signes profonds lexicaux (les lexies) contiennent une information qu'on trouve en principe dans le lexique d'un modèle Sens-Texte (sous l'appellation de tableau de régime), tandis que les signes profonds grammaticaux ont leur équivalent dans la partie grammaire du modèle Sens-Texte. On peut, si on le veut, restituer la partition lexique-grammaire en présentant comme lexique l'ensemble des signes dont le signifiant ou le signifié contient un lexème (c'est-à-dire les lexies et les lexes) et dire que la grammaire constitue le reste (les grammies, les grammes, les syntagmes et les morphes). Il reste que les deux ensembles de signes sont exprimés dans le même formalisme et se combinent suivant les mêmes principes, alors qu'en TST, lexique et grammaire sont traités très différemment : le lexique est une ressource statique, tandis les règles de grammaires sont des règles de correspondance, qui viennent dynamiquement interroger le lexique (voir [A2] pour une procédure de synthèse détaillée).

Nous allons maintenant présenter les différents modules de GUST : l'interface sémantique-syntaxe (les signes profonds), l'interface syntaxe-morphologie (les signes intermédiaires) et l'interface morphologie-phonologie (les signes de surface). Nous continuons à illustrer notre propos par l'exemple (2), dont nous avons donné les différentes représentations à la section précédente.

Interface sémantique-syntaxe de GUST

Nous avons donné, dans la Figure 7 de la section précédente, la structure élémentaire associée au signe PARLER. Cette structure indique que le signifié de PARLER est le sémantème 'parler' qui possède deux arguments et que son signifiant est le lexème PARLER qui possède

---

propage à tous les traits *t* de la chaîne. C'est ce qu'on appelle la percolation de traits. En un sens, on masque le fait que la structure qui a instancié le trait *t* en se combinant avec une structure à un bout de la chaîne s'est en fait combinée avec l'ensemble des structures de la chaîne, puisqu'elle a instancié la valeur du trait *t* dans l'ensemble des structures de la chaîne.

deux actants syntaxiques, un sujet (*suj*) et un complément d'objet indirect (*iobj*)[40]. Cette structure indique également la diathèse de PARLER, c'est-à-dire la correspondance entre les arguments sémantiques et les actants syntaxiques. Cette structure indique enfin que PARLER est un verbe (V) et qu'il doit se combiner avec trois grammies (= grammes profonds) de mode (m), de temps (t) et de voix (v). La flèche (→) devant un trait grammatical indique que ce trait doit être rempli. Nous appellerons *valence grammaticale* d'une lexie l'ensemble des traits grammaticaux à remplir. Nous distinguons clairement les traits grammaticaux de la valence grammaticale des traits catégoriels comme la partie du discours, le genre des noms, le groupe des verbes, etc. Suivant l'usage de la TST, les traits catégoriels sont placés entre parenthèses.

Une structure sémantico-syntaxique est *saturée* si tous les nœuds syntaxiques sont étiquetés par un lexème, si tous les nœuds sémantiques (figurés par un trait sém) sont étiquetés par un sémantème et si toutes les valences grammaticales sont saturées. Les variables *x*, *y*, … qui apparaissent après certains traits servent uniquement à indiquer que deux traits partagent la même valeur[41], elles n'instancient donc pas le trait après lequel elles figurent.

Nous allons maintenant présenter les structures représentant des grammies (= des signes grammaticaux profond). Les grammies sont déclenchées par la valence grammaticale d'une lexie[42] : une grammie ne peut se combiner avec une lexie que si le trait grammatical qu'elle instancie est à remplir. Nous proposons trois exemples de grammies dans la Figure 8 : le temps présent a pour signifié le sémantème '*présent*' et pour signifiant le grammème présent. Le sémantème '*présent*', comme tous les sémantèmes grammaticaux, est un prédicat unaire. Son argument est le sémantème lexical qui instancie le trait sém se trouvant sur le même nœud syntaxique ; le sémantème '*présent*' est lui-même présenté comme valeur du trait grammatical t qu'il instancie. Une grammie comme la voix active est un peu bizarre, puisqu'elle n'a ni un vrai signifié, ni un vrai signifiant : elle se contente de décharger le trait v (ce que nous indiquons par v: →)[43]. La grammie de l'indicatif est du même genre que la voix

---

[40] La structure de la Figure 7 indique que PARLER possède un sujet. Cette propriété, qui est commune à tous les verbes du français, devrait être exprimée à un plus haut degré de généralité. Une solution, adoptée par HPSG, consiste à définir une ontologie sur l'ensemble des signes et à faire hériter par tous les verbes la propriété d'avoir un sujet. Une alternative, défendue par l'école chomskyenne, est de considérer que la propriété pour une proposition d'avoir un sujet n'est pas une propriété lexicale du verbe et doit être totalement dissociée de la description du verbe, c'est-à-dire dans notre cas, donner un signe indépendant de la lexie verbale.

[41] HPSG utilise, pour indiquer le partage des valeurs entre traits, des numéros dans des boites : 1, 2, …

[42] Dans la version actuelle de GUST, les traits grammaticaux de mode, temps, voix sont tous les trois introduits par une lexie verbale. Peut-être serait-il préférable de ne pas faire introduire le temps directement par le verbe, mais plutôt par le mode, d'autant qu'à chaque mode correspond une catégorie flexionnelle du temps différente (deux temps à l'impératif contre au moins huit à l'indicatif).

[43] Il est d'usage de considérer les voix comme une catégorie flexionnelle et d'opposer à la voix passive une voix active. Il semble formellement plus simple de ne pas du tout considérer de voix active : la voix passive est alors un signe qui peut se combiner librement avec un verbe. Dans ce cas, la "forme active" résulte de la non combinaison du verbe avec une voix. Ceci permet de traiter élégamment les combinaisons de voix que l'on observe dans certaines langues comme l'indonésien (cf. [A1]) ou la combinaison du passif avec le causatif

active, si ce n'est qu'elle impose que le sémantème soit un nœud communicativement dominant (je n'entrerai pas plus dans les détails de ce point qui n'est pas formalisé pour l'instant ; pour la notion de nœud communicativement dominant, voir Polguère 1990 et [A2] ou [D1]).

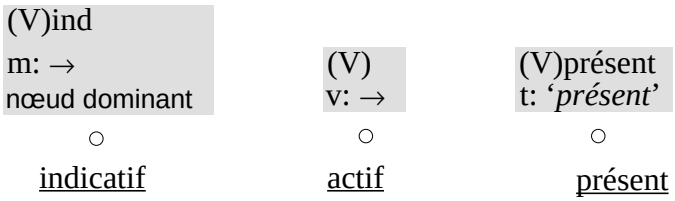


Figure 8 Les grammies de mode indicatif, voix active et temps présent

Nous allons maintenant présenter deux grammies un peu plus complexes, le passé composé et le passif, qui donnent tous les deux des formes verbales composées. La Figure 9 montre la combinaison du passé composé avec la lexie MANGER.

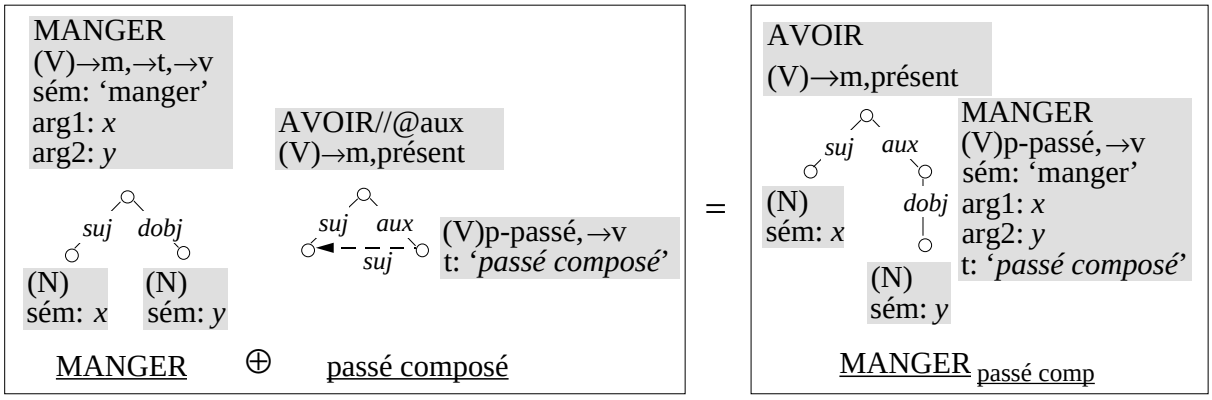


Figure 9 La combinaison de la lexie MANGER et de la grammie passé composé

La complexité du passé composé vient de l'introduction d'un auxiliaire[44] et du fait que certains éléments restent sur le participe passé, mais que d'autres sont entraînés par

comme en italien (*Il libro è stato fatto leggere agli studenti di Maria* 'Le livre a été fait lire aux étudiants par Marie'). Par contre, si le passif et le réfléchi n'appartiennent plus à une même catégorie flexionnelle, il faudra prendre quelques précaution dans la formalisation pour éviter qu'ils ne se combinent entre eux et donnent les formes agrammaticales **Un livre s'est acheté par Pierre* (forme passive du réfléchi *Pierre s'achète un livre*) ou **Jean s'est acheté par Pierre* (forme réfléchi du passif *Jean est acheté à Jean par Pierre*).

[44] La règle du passé composé, comme celle du passif plus loin, comporte une omission importante : la sémantique attachée à l'auxiliaire. Il est clair que l'auxiliaire n'a pas de sémantique propre et qu'il est un des éléments exprimant le sémantème '*passé composé*'. Néanmoins, on en peut attribuer à l'auxiliaire un trait ¬sém, car celui-ci peut être l'argument d'un autre élément (par exemple l'argument de *pense* comme dans *Jean pense que Pierre a mangé*) certains adverbes peuvent dépendre syntaxiquement de l'auxiliaire (comme par exemple la négation dans *Pierre n'a pas mangé*) et qu'il faut assurer que le sémantème du verbe participe passé soit bien l'argument sémantique de l'adverbe. Une solution est de donner à l'auxiliaire la même sémantique que celle du participe passé en partageant les valeurs des traits sém de l'auxiliaire et du participe passé.

l'auxiliaire, notamment le sujet du verbe[45]. Formellement, nous traitons ce changement de gouverneur grâce à un type particulier de lien que nous appelons une *quasi-dépendance* et que nous représentons par une flèche hachurée. Une quasi-dépendance peut seulement s'unifier avec une dépendance de même étiquette et le résultat de l'unification des deux liens est effacé. Contrairement à [D1] et [A3], nous ne gardons pas les quasi-dépendances dans la structure syntaxique du fait qu'elles ne jouent aucun rôle dans la bonne formation de la représentation syntaxique (notamment la structure d'arbre) et qu'elle ne sont pas considérées par l'interface syntaxe-morphologie. Nous pouvons ainsi imposer qu'une structure syntaxique bien formée ne contienne plus de quasi-dépendance, ce qui nous permet d'assurer que toutes les quasi-dépendances que nous introduisons se combineront bien avec des dépendances. Comme on le voit dans la Figure 9, la quasi-dépendance *sujet* de la structure du passé composé sert uniquement à assurer que le sujet de l'auxiliaire soit bien reconnu comme le premier argument de 'manger'. Notons également que la voix reste sur le participe passé, tandis que le mode est exprimé sur l'auxiliaire. Enfin, l'expression AVOIR//@aux signifie que l'auxiliaire est AVOIR à moins que le trait aux soit instancié : dans ce cas, l'auxiliaire est la valeur du trait aux, qui est alors ÊTRE (@ désigne l'opérateur qui donne la valeur d'un trait).

Le cas du passif est plus complexe que celui du passé composé : en plus de l'introduction d'un auxiliaire il y a une redistribution des fonctions syntaxiques. Qui plus est, le participe passé du passif peut être utilisé sans auxiliaire comme modifieur de nom (*le livre volé par Pierre*). Pour cette raison, nous ne traitons pas le passif comme un signe élémentaire, mais comme le composé de deux signes[46] : participe passif ⊕ copule (voir Figure 10).

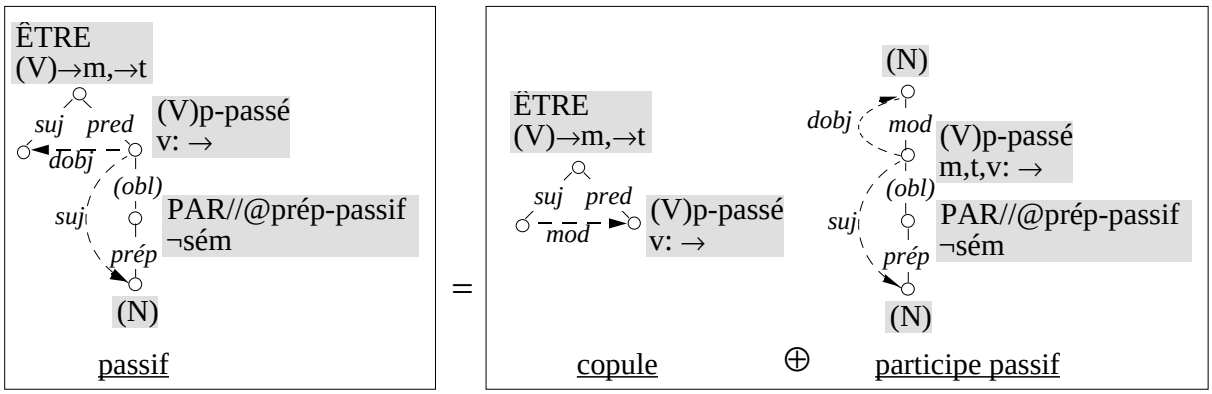


Figure 10 Le passif

La redistribution des fonctions pour le passif (et le participe passif) est encore assurée par des quasi-dépendances. De même que le passage du gouverneur du participe passif comme sujet

[45] Le fait que le sujet dépend syntaxiquement de l'auxiliaire ne fait aucun doute : c'est par rapport à l'auxiliaire que le sujet se place et c'est avec lui qu'a lieu l'accord.

[46] Ceci est un nouvel argument pour ne pas traiter le passif comme un élément de la catégorie flexionnelle de voix (voir Note 43).

de la copule (par l'intermédiaire de la quasi-dépendance *mod*). A noter que la structure de la copule est la même que celle utilisée pour les adjectifs (cf. [D1]).

Nous pouvons maintenant donner la dérivation (à l'étage sémantico-syntaxique) de la phrase (2) (Figure 11). Il est important de noter que la structure utilisée pour MANGER est exactement la même que celle qui serait utilisée pour dériver *Zoé mange la soupe* ou *La soupe est mangée par Zoé*. Contrairement à d'autres modèles, comme TAG ou HPSG, il n'est pas nécessaire de considérer une structure particulière pour la forme infinitive de MANGER. Qui plus est, il n'y a même pas de grammie infinitif qui intervienne dans la dérivation de (2). L'infinitif est imposé par ESSAYER et la mise sous forme infinitive de MANGER est entièrement assurée par ESSAYER, notamment la non réalisation du sujet de MANGER[47].

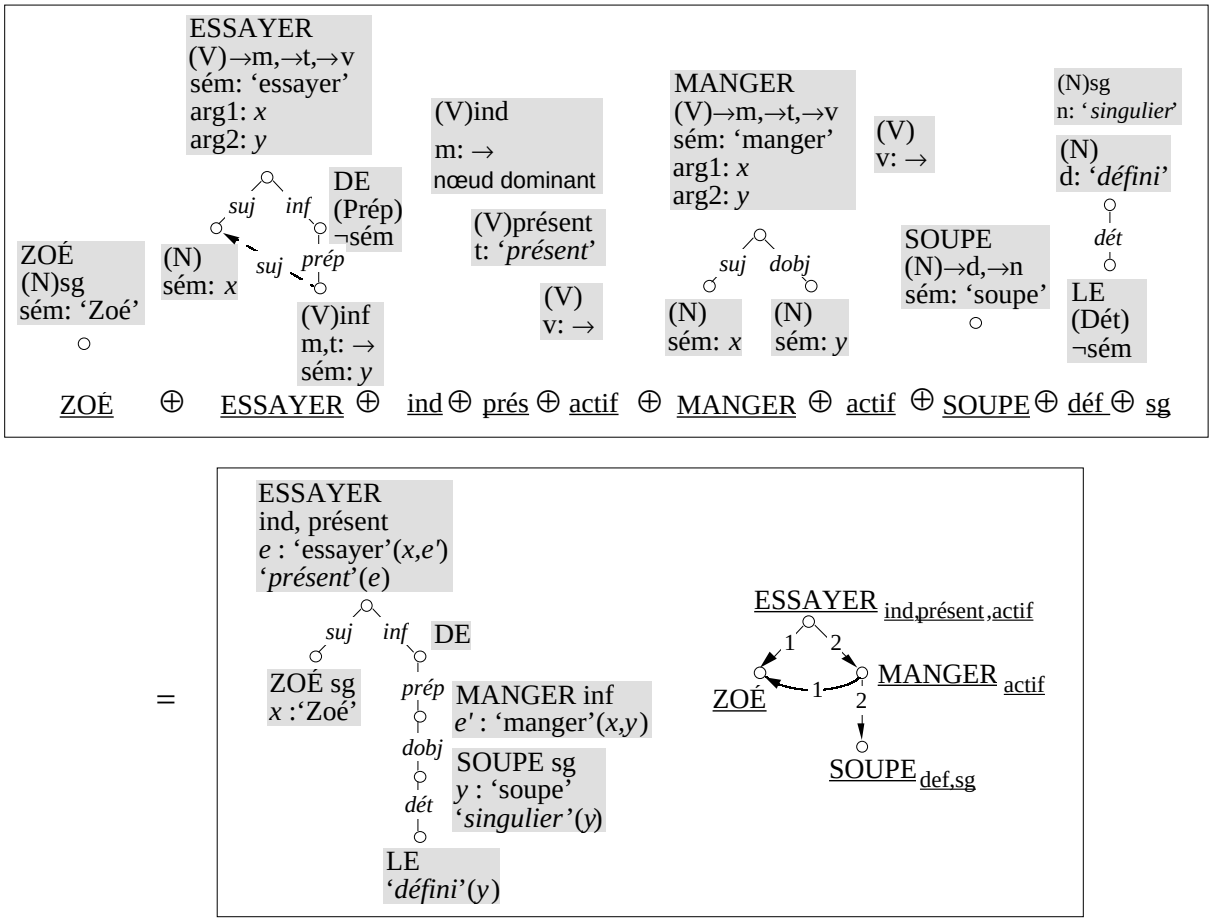


Figure 11 Dérivation de (2)

Nous allons maintenant discuter d'une notion importante, la *structure de dérivation*. La structure de dérivation est le témoin de la dérivation : elle décrit comment des règles (ici les signes profonds) se sont combinées pour dériver une structure. La notion d'arbre de

[47] Dans [D1], la question n'était pas encore tranchée et je suggérais qu'une forme infinitive de MANGER était utilisée dans le module sémantique, laquelle était obtenue par des règles lexicales.

dérivation[48] a été introduit par Vijay-Shanker 1987 pour les TAG. Différents travaux de Rambow & Joshi 1992 à [C6] ont montré que la structure de dérivation TAG s'apparente à une représentation linguistique sémantico-syntaxique de la phrase dérivée.

La structure de dérivation profonde (ou sémantico-syntaxique) de (2) est donnée en bas à droite de la Figure 11. Cette structure est un graphe orienté partiellement hiérarchisé, c'est-à-dire un graphe orienté sur lequel se superpose des relations hiérarchiques entre certains nœuds. La structure de dérivation est définie comme suit. Un arc du graphe correspond à une dépendance sémantique (= une relation prédicat-argument) et est étiqueté par le numéro d'argument comme la dépendance sémantique correspondante. Une dépendance sémantique qui correspond également à une dépendance syntaxique est hiérarchisée et représentée par une flèche droite, le gouverneur syntaxique étant placé plus haut comme dans un arbre syntaxique. Une dépendance sémantique qui ne correspond pas à une dépendance syntaxique est représentée par une flèche courbe. Une combinaison syntaxique qui ne correspond pas à une dépendance sémantique est représentée par une ligne droite non orientée et non numérotée (voir exemples Figure 13 et Figure 22). Les grammies font l'objet d'une convention spéciale et sont indiquées en indice de la lexie à laquelle elle s'applique.

La structure de dérivation sémantico-syntaxique d'une phrase rend compte des combinaisons entre les signes profonds de cette phrase. Avec les conventions adoptées, la structure de dérivation sémantico-syntaxique s'apparente à l'arbre de dépendance syntaxique profond de la TST. Ainsi, alors que nous avions rejeté l'arbre syntaxique profond comme niveau de représentation intermédiaire entre les représentations sémantique et syntaxique (de surface), une structure comparable est en fait sous-jacente à notre modélisation. Nous n'allons pas discuter des différences et similitudes entre notre structure de dérivation sémantico-syntaxique et l'arbre syntaxique profond de la TST. Notons simplement que ce parallèle permet, à mon avis, de mieux saisir l'intuition qui sous-tendait l'introduction du niveau syntaxique profond dans la TST et de mieux comprendre les nombreux problèmes de définition que pose ce niveau intermédiaire dans le cadre de la TST[49].

---

[48] On parle également d'arbre de dérivation pour une grammaire de réécriture hors-contexte (Chomsky 1957). Il ne s'agit pas exactement d'une structure de dérivation au sens considéré en TAG ou en GUST, bien que les deux notions soient reliées. Notons par ailleurs que les structures de dérivation de TAG ne sont pas à proprement des arbres. Par exemple, dans la dérivation usuelle de *little red books*, *red* s'adjoint sur le nœud N de *books*, puis *little* s'adjoint à son tour sur un nœud N. Ce dernier nœud résulte du nœud N de *books* et de la racine de *red* et en un sens *little* se combine simultanément avec les structures de *books* et *red*. Dans un soucis d'avoir un arbre (?!) comme structure de dérivation, un seul lien est retenu : Vijay-Shanker 1987 retient le lien entre *little* et *red*, tandis que Shieber & Schabes 1994 ont proposé de retenir le lien entre *little* et *books*, plus intéressant du point de vue de l'interprétation de la structure de dérivation.

[49] Alors que les dépendances sémantiques et les dépendances syntaxiques de surface sont assez bien caractérisées, les dépendances syntaxiques profondes ne sont jamais clairement caractérisées. De plus, celles-ci sont toujours définies dans un deuxième temps, après les dépendances sémantiques et syntaxiques de surface (cf. Mel'čuk 1988).

Nous allons terminer cette section en présentant encore quelques exemples de signes profonds. Parmi les signes profonds lexicaux, il faut noter deux familles importantes : les locutions (voir Figure 12) et les fonctions lexicales (voir Section 6). Le fait de distinguer trois étages de signes rend le traitement des locutions particulièrement économique : le signe profond associé à une locution n'introduit que l'information spécifique à la locution. Tout ce qui concerne les mots de la locution est mis en facteur aux étages intermédiaires et de surface avec les emplois usuels de ces mots, c'est-à-dire leur emploi en dehors de la locution. Ceci n'est pas le cas dans le traitement des autres modèles, que ce soit HPSG, TAG ou LFG. Par contre un traitement assez similaire est proposé dans le cadre de la TST (Mel'čuk 1995b). Concernant les deux locutions de la Figure 12, on notera que seule la racine syntaxique porte un trait sém ; le fait d'attribuer un trait ¬sém aux autres nœuds permet de bloquer toute modification sur ces nœuds (voir [D1] pour une illustration avec la locution LA MOUTARDE MONTER AU NEZ).

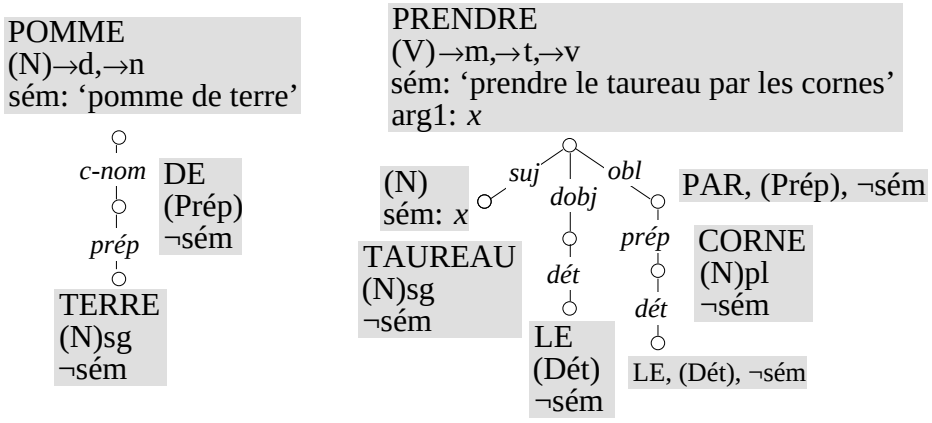


Figure 12 Les locutions POMME DE TERRE
et PRENDRE LE TAUREAU PAR LES CORNES

Nous allons terminer par un exemple assez intéressant de signe profond, lequel n'est à proprement parler ni une lexie, ni une grammie. Commençons par revenir sur l'infinitif. Nous avons remarqué, dans la dérivation de (2), que l'infinitif sur le verbe subordonné était attribué par le verbe recteur et ne résultait donc pas de la combinaison avec une grammie infinitif. Par contre, il est nécessaire de considérer une grammie infinitif dans l'analyse d'une construction comme *Voler est un délit*. La structure associée à l'infinitif apparaît dans la Figure 13 ci-dessous. Cette structure supprime le sujet potentiel du verbe, mais attribue le sémantème '*générique*' à l'argument du verbe qui est potentiellement sujet[50]. Attention, ce n'est pas forcément le 1er actant sémantique du verbe qui devient générique, c'est celui qui s'exprimerait comme sujet s'il n'y avait pas d'infinitif : par exemple, dans *Etre volé est un malheur*, c'est le 2ème actant du verbe qui est générique. Il est donc bien nécessaire de mentionner la fonction sujet dans la structure élémentaire associée à la grammie infinitif.

[50] Encore une fois, nous contournons les questions de sémantique : nous attribuons au sujet générique du verbe l'infinitif le sémantème '*générique*' sans chercher à caractériser davantage ce qu'est réellement ce sens.

Venons-en au signe profond qui nous intéresse, celui qui introduit la préposition À dans la construction (6) :

(6) *un livre facile à lire*

La construction illustrée en (6) exprime le sens ‘un livre tel que lire ce livre est facile’ et pallie l’impossibilité de construire une relative (**un livre que lire _ est facile*)[51]. Dans cette construction, j’estime que les lexies FACILE[52], LIRE et LIVRE sont les mêmes que dans *lire ce livre est facile*[53]. Je propose donc un signe, que j’appelle Adj À Vinf, qui effectue un changement de diathèse de l’adjectif et s’apparente à une voix adjectivale (voir Figure 13). En même temps, ce signe n’a pas les caractéristiques morphologiques d’une grammie et il serait exagéré de parler réellement d’une voix. Par ailleurs ce signe n’introduit que le lexème À qui est sémantiquement vide. Ce signe correspond donc avant tout à une construction syntaxique particulière[54].

[51] Certains auteurs (par exemple Kayne 1974-75) proposent de dériver *ce livre est facile à lire* de *il est facile de lire ce livre* par montée du sujet. Cette dernière construction est la forme impersonnelle de *lire ce livre est facile*. Toutes ces constructions sont effectivement liées, mais, même dans une perspective transformationnelle, il me paraît préférable de lier directement (6) à *lire ce livre est facile*, plutôt que de passer par de tels intermédiaires.

[52] Tel qu’est décrit FACILE dans la Figure 13, il exige que son 1er argument sémantique soit un nom. En fait ce 1er argument peut être aussi un verbe infinitif sous certaines conditions. Tesnière considère que le verbe infinitif est un verbe translaté en nom, c’est-à-dire qu’il possède plus ou moins la distribution d’un nom et que donc il se comporte comme un nom vis-à-vis de son gouverneur tout en continuant à se comporter comme un verbe vis-à-vis de ses dépendants. Ainsi chaque mot a comme deux faces avec deux parties du discours potentiellement différentes. La contrainte exigeant que le 1er argument sémantique de FACILE soit un nom ne s’applique en fait qu’à l’une des deux faces de cet argument, la face inférieure lorsque l’adjectif modifie son argument (comme dans *un livre facile*) et la face supérieure lorsque cet argument est dépendant (comme dans *ce livre est facile*). Pour cette raison, un verbe infinitif (qui apparaît comme un nom vis-à-vis de son gouverneur) peut être le 1er argument d’un adjectif lorsqu’il est dépendant (*lire ce livre est facile*), mais pas lorsqu’il est gouverneur (**il peut lire facile ce livre*). Pour formaliser cela, il nous suffirait distinguer pour chaque mot deux parties du discours, un partie du discours supérieure pour la combinaison avec le gouverneur et une partie du discours inférieure pour la combinaison avec les dépendants (voir [A7] pour une formalisation complète en HPSG).

[53] Dans [D1], ce problème n’est pas encore résolu et deux structures différentes sont considérées pour les deux emplois de *facile*.

[54] Le fait que certaines informations correspondent à des constructions, plutôt qu’à des lexies en particulier est un argument utilisé par certaines approches comme HPSG pour maintenir des règles syntagmatiques (Ginzburg & Sag 2000). On voit sur cet exemple qu’on peut considérer des signes associés à des constructions sans pour autant considérer des syntagmes. On peut également mentionner certains signes profonds comme l’expression/construction illustrée par : *pour être con, il est con* ; *pour être nul, c’est nul* ; *pour être mal barré, on est mal barré*. Ce type d’expression (étudié et décrit en terme de signe linguistique dans Mel’čuk 1995a) n’est pas vraiment une locution dans le sens où les seuls marqueurs sont des mots vides (POUR ÊTRE …, … ÊTRE …) comme dans la construction Adj À Vinf.

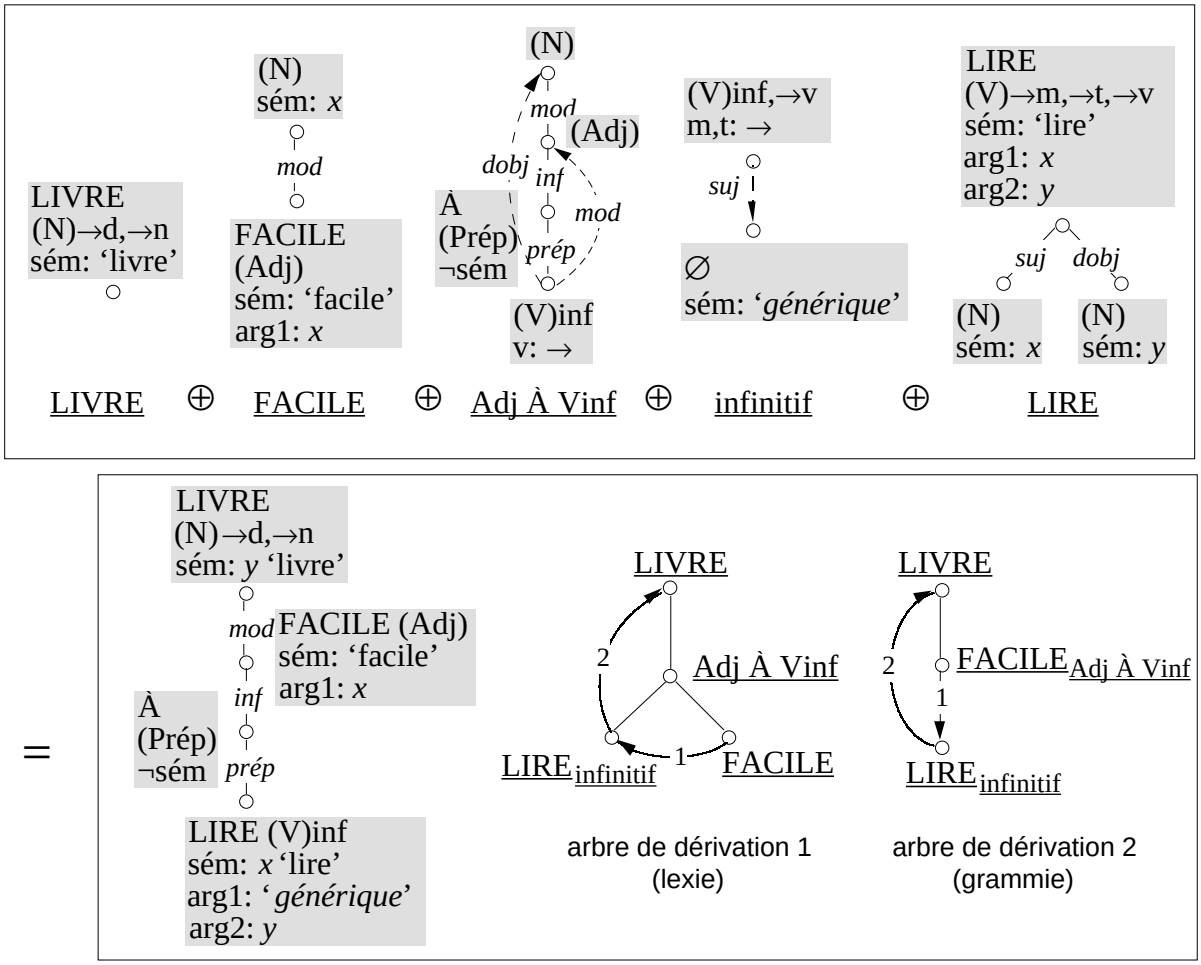


Figure 13 Dérivation de (6)

On notera que la structure de Adj À Vinf se combine avec trois autres structures. Dans le premier arbre de dérivation de la Figure 13, nous avons traité Adj À Vinf comme une lexie et nous lui avons donc attribué un nœud, mettant en évidence les combinaisons de Adj À Vinf avec ses trois "arguments". L'ordre de ces combinaisons est complètement libre. Adj À Vinf peut même se combiner avec l'infinitif avant de se combiner avec le verbe LIRE. Le seul ordre de combinaison qui est impossible est que l'adjectif FACILE se combine avec LIRE avant de se combiner avec Adj À Vinf. Ceci peut nous inciter à traiter Adj À Vinf comme une grammie qui s'applique à FACILE. Nous obtenons alors le deuxième arbre de dérivation de la Figure 13.

Nous arrêterons là la présentation du module sémantique de GUST. J'espère avoir convaincu le lecteur que le point de vue théorique et formel adopté est d'une grande économie et qu'une même structure permet de couvrir tous les emplois d'un signe profond, quelque soit la construction dans laquelle il figure et quelle que soit sa forme morphologique, c'est-à-dire quels que soit les signes grammaticaux qui se combinent avec lui. La démonstration ne pourra être parfaite que quand un fragment plus conséquent du modèle aura été développé. Parmi les points qui n'ont pas encore été suffisamment étudiés figurent le traitement de la translation évoqué précédemment et le traitement des différentes sous-catégorisations d'une lexie (par

exemple, *demander* N *à* N, *à* N *de* V_{inf}, *que* V_{subj}, *à* V_{inf}, …). Les deux points sont liés puisque la variabilité des sous-catégorisations est justement liée aux possibilités de translation.

Interface syntaxe-morphologie de GUST

En l'absence d'un réel traitement de la morphologie, l'interface syntaxe-morphologie actuelle de GUST traite essentiellement l'ordre des mots, l'accord et le régime. Les signes intermédiaires sont de trois types : les lexes, les grammes et les syntagmes. Les lexes et les grammes qui correspondent à une grammie font l'objet d'un traitement trivial, c'est-à-dire qu'il y a une correspondance biunivoque entre leur signifié et leur signifiant, qui autorise à ne pas les distinguer et donc à ne même pas mentionner ces signes dans l'architecture globale du système. La seule chose que font les lexes est d'introduire une nouvelle valence grammaticale : ainsi, en français, les verbes requièrent-ils des grammes d'accord de personne et de nombre (→p, →n), les adjectifs des grammes d'accord de nombre et de genre (→n, →g) et les pronoms personnels des grammes de cas (→c). Les structures des grammes d'accord du verbe et le gramme **nominatif** sont donnés Figure 14. En fait, la première structure est un patron commun à deux signes suivant que la variable *n* correspond au grammème singulier (sg) ou pluriel (pl).

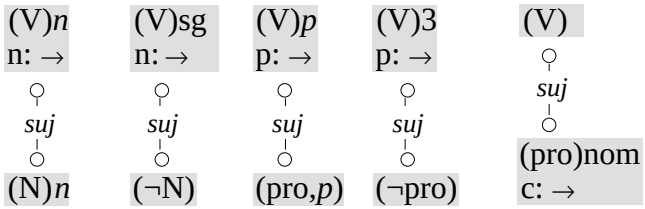


Figure 14 Les grammes d'accord du verbe et le **nominatif**

Comme on le voit, les grammes d'accord et les grammes de régime sont contrôlés par les fonctions syntaxiques : en français, le verbe s'accorde avec son *sujet* et avec son *objet direct* dans certains cas (accord du **participe passé** gouverné par **AVOIR** avec un *objet direct* qui le précède) ; en allemand, l'adjectif *épithète* s'accorde avec le nom, mais l'adjectif *attribut* est invariable, etc. Les fonctions syntaxiques contrôlent également les redistributions comme le passif ou Adj À Vinf, ainsi que la pronominalisation[55]. Enfin, les fonctions syntaxiques contrôlent l'ordre des mots, c'est-à-dire les syntagmes. La fonction syntaxique associée à une dépendance est déterminée par l'ensemble des signes qui peuvent s'appliquer sur cette dépendance : deux dépendances recevront la même fonction syntaxique si elle se comporte de la même façon vis-à-vis de l'accord, du régime, des redistributions, de la pronominalisation et de l'ordre des mots, c'est–à-dire vis-à-vis de tous les phénomènes qui discriminent les dépendances (Mel'čuk 1988, Iordanskaja & Mel'čuk *à paraître*).

[55] Cf. Blanche-Benveniste 1975 pour une caractérisation des positions syntaxiques basée sur la distribution des pronoms.

Nous allons maintenant proposer une première version des syntagmes. Une deuxième version plus élaborée sera développée dans la Section 5. Un syntagme associe une dépendance à une spécification d'ordre et plus précisément une fonction syntaxique à une distance entre deux mots-formes. Nous indiquons la distance comme valeur du trait pos. La distance permet de placer les co-dépendants d'un nœud les uns par rapport aux autres : plus un nœud a une valeur de pos importante (en valeur absolue) plus il est éloigné de son gouverneur.

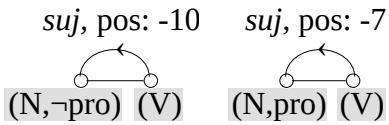


Figure 15 Les syntagmes ***sujet_canonique*** et ***sujet_clitique***

Comme les autres signes, les syntagmes se combinent par unification. On impose que le résultat de la combinaison des syntagmes soit un arbre ordonné projectif (voir [D1] ou [A3] pour la définition de la projectivité). On peut également interpréter la distance comme une assignation de champs : chaque lexe ouvre une boîte avec un certain nombre de champs numérotés par exemple de –20 à +20. Le lexe occupe le champ 0 et chaque dépendant est placé dans un champ. On peut préciser pour chaque champ si celui-ci peut accueillir un ou plusieurs éléments et s'il peut rester vide ou non. Si on impose que chaque dépendant doit se placer dans la boîte ouverte, on assure la projectivité. Si on veut traiter des constructions non projectives, on peut autoriser un dépendant à sortir de la boîte de son gouverneur en contrôlant par exemple les frontières de boîtes traversées par ce dépendant. Le formalisme que nous venons de décrire est celui proposé dans [C16] pour traiter l'ordre des mots en allemand. Dans le cas de l'allemand, plutôt que de numéroter les champs comme ici nous avons repris la terminologie usuelle du modèle topologique de l'allemand (Drach 1937, Bech 1955). Nous reprendrons cette discussion dans la Section 5.

Interface morphologie-phonologie de GUST

Comme nous l'avons dit, GUST ne traite pas réellement la morphologie pour l'instant et associe directement un lexème accompagné d'une liste adéquate de grammème et une chaîne phonologique ou une chaîne graphique

MANGER
(V)ind,présent,1,pl
graph: *mangeons*
phon: /mãjõ/

Figure 16 Le signe de surface *MANGEONS*

3.4 Analyse et synthèse en GUST

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 une représentation phonologique ou graphique. 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).

Nous présentons dans la Figure 17 la dérivation complète de (2). Cette dérivation met en jeu 4 lexies (ZOÉ, ESSAYER, MANGER, SOUPE), 6 grammies, 5 syntagmes (correspondant aux 5 dépendances syntaxiques), 4 grammes d'accord et 6 signes de surface (correspondant aux 6 mots-formes). L'ensemble de ces signes peuvent être combinés entre eux par unification. L'ordre de combinaison des signes n'est pas complètement libre. Par exemple, on ne peut pas déclencher le gramme d'accord en nombre du déterminant avec le nom avant d'avoir combiner le nom avec une grammie de nombre. Certaines contraintes doivent être ajoutées, soit sur l'ordre de combinaisons des signes, soit sur les signes eux-mêmes, pour éviter qu'une dépendance se combine avec un syntagme avant d'être éliminé par une quasi-dépendance : par exemple, la relation *sujet* de MANGER ne doit pas se combiner avec un syntagme, puisqu'elle va disparaître lors de la combinaison de MANGER avec ESSAYER.

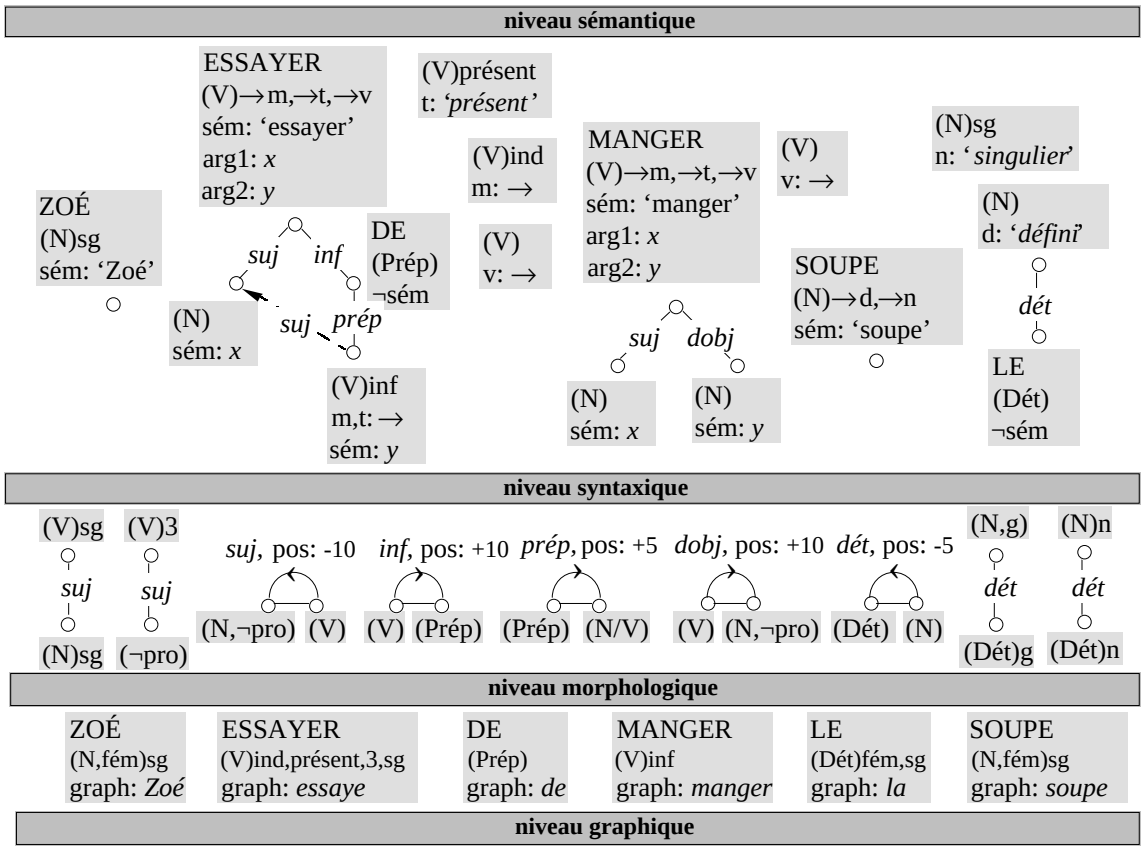


Figure 17 Dérivation de (2)

Quelques mots de comparaison avec TAG. Une dérivation de (2) en TAG mettrait en jeu 5 arbres de dérivations (correspondant à ZOÉ, ESSAYER_DE, MANGER, LE et SOUPE), plus 6 règles de lemmatisation équivalentes à nos 6 signes de surface. A part la plus grande

articulation de GUST, la dérivation en TAG et en GUST suivent des principes similaires. De même qu'une grammaire TAG, une grammaire GUST peut aisément être utilisée en analyse comme en synthèse. De plus, contrairement à TAG, chaque signe est associé à une seule structure élémentaire (alors qu'en TAG les formes verbales, même non ambigues, sont associées à des centaines de structures élémentaires), ce qui évite l'introduction d'un grand nombre d'ambiguïtés artificielles au cours du processus d'analyse. Un système de génération basé sur TAG comme G-TAG (Danlos 1998) peut facilement être adapté à GUST. La aussi le gain est notable. Par exemple, si l'on prend l'exemple (2), lors du choix de la structure pour le sémantème 'manger', sachant qu'un verbe est requis, une fois qu'on a choisi d'exprimer 'manger' par le verbe MANGER, on déclenche la lexie MANGER, alors qu'avec TAG, il faut choisir parmi les multiples structures associée à MANGER en tenant compte du fait que le verbe est à l'infinitif et que son objet n'est ni cliticisé, ni extrait (ou bien on ne choisit pas et on les essaye toutes !).

Comme nous l'expliquons dans [D1], en raison de l'articulation du modèle, plusieurs stratégies d'analyse ou de synthèse sont possibles. Les deux extrêmes sont les stratégies horizontale et verticale.

- Le principe de la *stratégie horizontale* est de déclencher les signes étage par étage. En analyse, on commencera donc par déclencher les signes de surface (lemmatisation), puis les signes intermédiaires (shallow parsing), puis les signes profonds. En synthèse, on fera l'inverse. La difficulté de la stratégie horizontale est de pouvoir gérer l'ambiguïté à chaque niveau de représentation intermédiaire, puisqu'on ne peut pas et on ne doit pas chercher à désambiguïser avant d'atteindre la représentation finale.
- Le principe de la *stratégie verticale* est de déclencher à partir d'un signe les signes des différents étages qui correspondent au même mot. Par exemple, en analyse, à partir de *essaye*, on déclenchera le signe de surface *ESSAYE*, puis la lexie ESSAYER et les grammies requises et enfin les syntagmes correspondant aux relations syntaxiques introduites par ESSAYER.

Différentes stratégies intermédiaires sont possibles. Par exemple, la synthèse incrémentale : le principe est de produire l'arbre syntaxique de haut en bas et les mots-formes dans l'ordre linéaire. Prenons l'exemple de (2). On commencera par déclencher le signe ESSAYER qui correspond au sommet syntaxique de la phrase. Ensuite, on déclenchera les grammies qu'il requiert, puis plutôt que de continuer à déclencher des signes profonds, on cherchera le premier mot de la phrase en déclenchant les syntagmes qui place un dépendant à gauche. Ceci va nous amener à déclencher le syntagme *sujet*, puis à déclencher toutes les règles qui permettent de produire *Zoé essaye* sans avoir commencé à traiter le reste de la phrase.

Une variante de la stratégie verticale consiste à lexicaliser le modèle, c'est-à-dire à précompiler la grammaire en regroupant par paquet tous les ensembles de règles qui peuvent être déclenchées par un même mot. Je voudrais compléter ici le chapitre consacré à la lexicalisation dans [D1]. Il est important de comprendre que la lexicalisation du modèle n'est pas un procédé immédiat et que des choix théoriques fondamentaux s'offrent à nous. D'abord,

il faut déterminer ce que l'on entend par mot, c'est-à-dire quelles seront les unités de la phrase qui déclencheront les règles lexicalisées. Ces unités sont a priori les lexies de la phrase, c'est-à-dire les unités sémantiques lexicales. Que fait-on des grammies qui donnent des mots-formes comme les verbes auxiliaires ou les articles ? Que fait-on des lexies qui ont tendance à s'affixer comme les clitiques ? Comme nous l'avons déjà dit, les réponses à ces questions ne sont jamais entièrement satisfaisantes et rarement consensuelles. Une fois déterminées les unités lexicales, il faut répartir les règles entre ces unités. Là encore, les choix ne sont pas immédiats. Prenons le cas de *mangent* $\text{MANGER}_{\text{ind,prés,3,pl}}$. Une règle lexicalisée pour cette forme est une combinaison de tous les signes qui peuvent être déclenchés à partir de cette forme. Une partie des signes sont déclenchés sans choix possibles[56] : MANGER, les grammies de mode, temps, voix et les grammes d'accord. Par contre, MANGER introduit des dépendances syntaxiques *sujet* et *objet direct* qui peuvent appeler plusieurs syntagmes (= règles d'ordre). De plus, on peut comme en TAG ou en HPSG, prendre en compte le fait qu'un des dépendants est extrait (ce n'est pas nécessaire en GUST). Avec le croisement des différents choix possibles, on obtient, après combinaisons des signes, plusieurs règles lexicalisées associées à *mangent* $\text{MANGER}_{\text{ind,prés,3,pl}}$. Cette stratégie, que nous appellerons la *lexicalisation par croisement*, nous donne une grammaire complètement lexicalisée à la TAG (avec tout de même moins de règles lexicalisées qu'en TAG). Nous donnons une des règles lexicalisées obtenues pour *mangent* $\text{MANGER}_{\text{ind,prés,3,pl}}$ par cette stratégie dans la Figure 18.

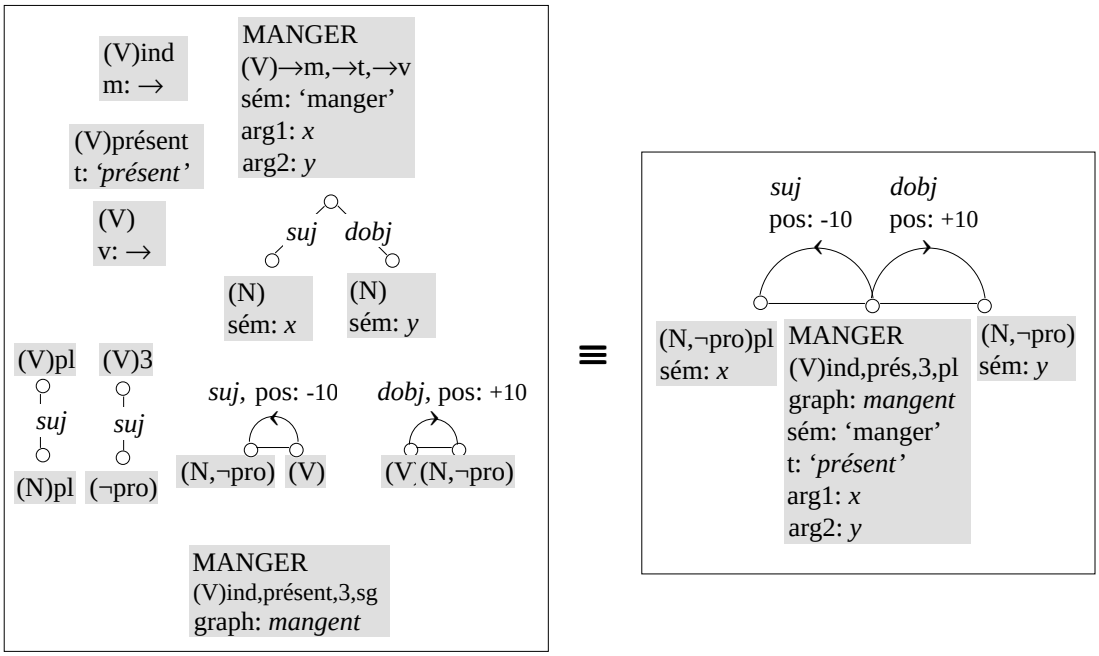


Figure 18 Lexicalisation par croisement

[56] Un problème vient également des différentes sous-catégorisations possibles du lexème concerné. Par exemple, l'*objet direct* de MANGER peut être omis. Doit-on considérer des signes distincts ou bien un seul signe avec un argument optionnel ?

Il est possible d'obtenir une grammaire complètement lexicalisée plus élégante en modifiant la stratégie de lexicalisation. Plutôt que d'appliquer aux dépendances *sujet* et *objet direct* de MANGER tous les syntagmes possibles, on ne leur applique que les syntagmes canoniques. Nous appelons cette stratégie la *lexicalisation par défaut*. On obtient une seule règle lexicalisée pour *mangent* $\text{MANGER}_{\text{ind,prés,3,pl}}$. Dans ce cas, les compléments qui n'occupent pas une position canonique, comme les clitiques, doivent modifier la règle d'ordre de la règle. Ceci peut être fait en attribuant un trait pos prioritaire au clitique (indiqué ici par un soulignement) ou en supprimant la dépendance par unification avec une quasi-dépendance et en introduisant à la place la dépendance appropriée (voir [A3] pour les détails). Nous donnons, dans la Figure 19, l'exemple de la combinaison de la règle par défaut de *mangent* $\text{MANGER}_{\text{ind,prés,3,pl}}$ avec celle du clitique *le* LUI_{acc}. A noter que la règle par défaut de *mangent* $\text{MANGER}_{\text{ind,prés,3,pl}}$ de la Figure 19 n'est pas la même règle que celle de la Figure 18 : contrairement à la règle de la Figure 18, cette règle doit permettre à un pronom personnel d'occuper les positions *sujet* et *objet direct* et n'a donc pas de restriction ¬pro.

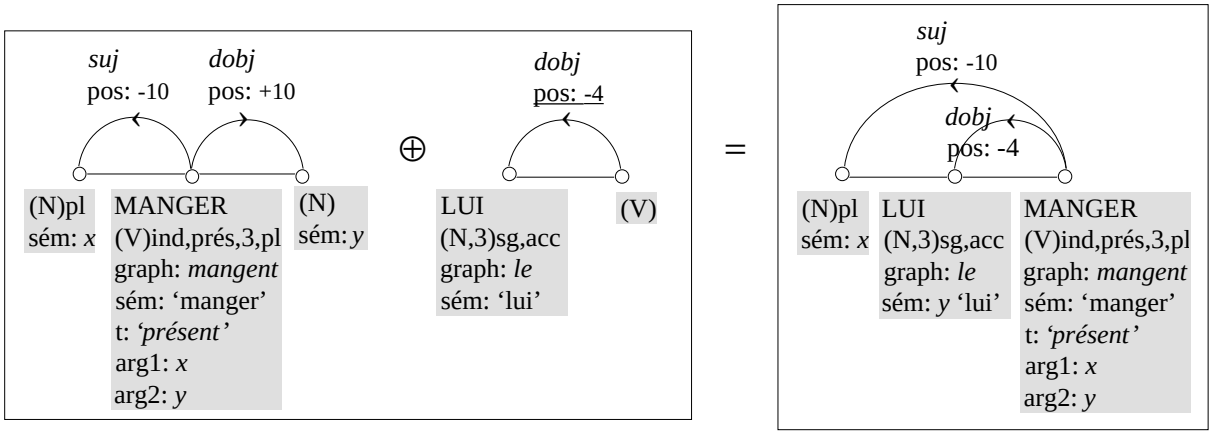


Figure 19 Lexicalisation par défaut

Comme on le voit, le formalisme de GUST permet d'écrire aussi bien des grammaires articulées que des grammaires lexicalisées. J'aimerais également insister sur le fait qu'un modèle articulé ne s'utilise pas nécessairement étage après étage et que de nombreuses procédures sont possibles. En particulier, un modèle articulé peut être utilisé de façon similaire à une grammaire lexicalisée sans qu'on doive pour autant pré-lexicaliser la grammaire (sachant qu'une grammaire lexicalisée, même optimisée comme ici, présente forcément des redondances et occupe donc davantage d'espace en mémoire).

Nous allons poursuivre notre présentation du formalisme de GUST en présentant trois problèmes linguistiques : l'extraction (Section 4), l'ordre des mots en allemand (Section 5) et les collocations (Section 6).

4 Extraction

On regroupe sous ce terme une série de phénomènes qui ont en commun d'avoir un groupe dit "extrait" (en gras dans les exemples qui suivent) : la relativisation (*la personne **à qui** Pierre*

parle), les interrogations directes et indirectes (***A qui*** *Pierre parle-t-il ?* ; *Je sais* ***à qui*** *Pierre parle*), le clivage (*C'est* ***à cette personne*** *que Pierre parle*), etc. Les phénomènes d'extraction partagent des propriétés distributionnelles (ce que l'on peut "extraire") et formelles, notamment l'usage des mots *qu-* (pronoms relatifs, interrogatifs, etc.). L'extraction pose des problèmes de modélisation qui en fait la pierre d'achoppement de tous les modèles syntaxiques depuis les années 70 (GB, GPSG, TAG, HPSG, LFG, CG, …).

Dans ce document de synthèse, je ne vais pas rentrer dans les détails de mes nombreux travaux touchant à l'extraction, mais en faire ressortir les idées principales. Dans la Section 4.1, je présente l'analyse de Tesnière sur la position des mots *qu-* qui est à la base des différentes modélisations formelles que j'ai développé, notamment des données assez énigmatiques connues sous le nom de "relatives infinitives". Dans la Section 4.2, je présente une notion fondamentale en linguistique, le nucléus verbal et nominal, qui permet de modéliser simplement l'apparente complexité des phénomènes d'extraction et notamment le caractère "non borné" de l'extraction. La Section 4.3 est consacrée à la modélisation de l'extraction en GUST et la Section 4.4 à des modélisations originales en TAG et en HPSG.

4.1 Position syntaxique et mots qu-

En grammaire syntagmatique comme en grammaire de dépendance, on considère généralement que chaque mot occupe une seule position syntaxique identifiée à un nœud de l'arbre syntaxique. Les mots *qu-* tels que pronoms relatifs et pronoms interrogatifs (dans les interrogatives indirectes) posent un problème intéressant, car ils remplissent à la fois un rôle pronominal et un rôle de marquage de la proposition, permettant ainsi à la proposition d'occuper des positions adjectivales pour les relatives et nominales pour les interrogatives indirectes. Tesnière (1959 : 561) propose d'attribuer au mot *qu-* deux positions syntaxiques, une position à l'intérieur de la proposition en tant que pronom et une position comme tête de la proposition en tant que translatif du verbe principal (voir ci-dessous). J'ai défendu cette hypothèse dans [A5] en l'étayant par l'étude de phénomènes syntaxiques qui ont été dégagés après les travaux de Tesnière, comme les coordinations de mots *qu-* (*Je me demande* ***où et quand*** *tu partiras*), l'alternance *qui-que* (*la personne* ***que*** *je pense* ***qui*** *dort*) ou les dites "relatives infinitives" (*Pierre cherche un endroit* ***où dormir***). Ce traitement des mots *qu-* est à la base des différentes modélisations de l'extraction proposées ci-après.

Avant de poursuivre le traitement de l'extraction, j'aimerais étayer l'analyse des mots *qu-* comme translatif en présentant le cas des interrogatives et des relatives infinitives ([A5]). La théorie de la translation de Tesnière (voir [D1] pour une présentation) repose sur l'idée que les positions syntaxiques requièrent des parties du discours précises : par exemple, les actants d'un verbe doivent être des noms et les dépendants d'un nom doivent être des adjectifs. Un élément d'une partie du discours différente peut occuper une telle position, mais doit pour cela être *translaté*. Ainsi pour occuper la position actancielle d'un verbe, un verbe doit être translaté en nom : les translatifs de verbe en nom sont l'infinitif (*Pierre sait li****re***), les conjonctions de subordination *que* et *si* (*Pierre sait* ***que/si*** *Marie vient*) et les pronoms interrogatifs (*Pierre sait* ***qui*** *vient*). Pour modifier un nom, un verbe doit être translaté en

adjectif : les translatifs de verbe en adjectif sont les participes (*le livre vol**é** par Pierre, la personne regard**ant** un livre*) et les pronoms relatifs (*le livre **que** Pierre regarde, la personne à **qui** Pierre parle*). Les interrogatives et les relatives infinitives posent problème à cette analyse des mots *qu-* comme translatifs de verbe. Dans une interrogative infinitive (*Pierre ne sait pas **qui** appel**er**, **où** dorm**ir***), on a deux translatifs de verbe en nom : l'infinitif et le mot *qu-*. On peut admettre que les translatifs opèrent conjointement, l'un sur l'autre. Cette hypothèse est appuyée par le fait que les interrogatives à verbe infinitif montrent une plus grande cohésion que les interrogatives à verbe fini, par exemple vis-à-vis de l'extraction hors de l'interrogative : *le livre que Pierre ne sait pas avec qui échanger* vs. ?**le livre que Pierre ne sait pas avec qui Marie échangera*. Le cas des relatives infinitives (*Je cherche un endroit **où** dorm**ir**, une personne à **qui** parl**er***) est beaucoup plus problématique : ici, s'il s'agit s'une relative, le mot *qu-* assure la translation du verbe en adjectif, tandis que l'infinitif assure la translation du verbe en nom. Les deux translatifs sont incompatibles : ils ne peuvent ni agir conjointement (comme dans les interrogatives infinitives), ni agir l'un après l'autre. L'hypothèse de Tesnière est mise en défaut … à moins que ce ne soit une autre hypothèse qui soit en défaut : le fait qu'il s'agisse ici de relatives. Il se trouve que les "relatives infinitives" ont un comportement bien énigmatique. Pour commencer, les positions où un nom prétendument modifié par une telle relative est accepté sont peu nombreuses (*Pierre cherche une personne à qui parler* vs. **Pierre fréquente une personne à qui parler* vs. **Les personnes à qui parler sont rares*). Or un modifieur, par définition, ne change pas la distribution de l'élément auquel il s'ajoute. Ceci laisse à penser que la "relative infinitive" est davantage une co-tête du nom qu'un modifieur. Par ailleurs, seules sont possibles les "relatives infinitives" dont le mot *qu-* est également possible dans les intégratives (ou "relatives sans antécédent")[57] : *Je cherche *un livre que lire*, **une personne dont parler*, ?**un moment où entrer* (alors que *un endroit où aller* est possible car le *où* locatif est à la fois relatif et intégratif). Ces deux observations convergent avec l'analyse de Tesnière et suggère que les "relatives infinitives" sont tout autant en position nominale qu'adjectivale et sont tout autant des intégratives que des relatives (si elles étaient vraiment des intégratives, on aurait **Je cherche une personne qui rencontrer*). Ce curieux phénomène d'entre deux, s'il est avéré, confirme une hypothèse défendue par exemple par Le Goffic 1993 qui veut que les relatives se soient formées à partir des intégratives par juxtaposition d'une intégrative avec un nom et recatégorisation de l'intégrative.

[57] Le terme de *relative sans antécédent* devrait être proscrit comme le recommande Le Goffic 1993. Ces propositions, souvent analysées comme des relatives dont l'antécédent aurait été effacé, ont moins de liens avec les relatives qu'avec les interrogatives indirectes (voir [A5] pour une argumentation). En particulier, les mots *qu-* utilisés dans les intégratives sont des mots pleins comme les pronoms interrogatifs avec une sémantique comparable et ils se distinguent nettement des pronoms relatifs, qui ont uniquement un rôle anaphorique. Ils se distinguent aussi par la forme : par exemple, le pronom intégratif objet est le pronom *qui* à valeur humaine (*je fréquente qui je veux*), tandis que le pronom relatif objet est *que* (*la personne que je veux fréquenter*).

4.2 Nucléus et dépendances non bornées

On appelle phénomène de *dépendance non bornée* le fait que deux éléments liés l'un à l'autre se trouvent à une distance structurelle (en termes de dépendances syntaxiques) potentiellement illimitée. L'*extraction* est probablement le plus étudié de ces phénomènes. Elle met en évidence deux phénomènes de dépendance non bornées. Le premier est que la distance entre le groupe extrait et son gouverneur syntaxique est potentiellement illimitée : *la personne à qui Pierre **parle**, à qui Pierre **a l'intention de parler**, à qui Marie **pense que** Pierre **a l'intention de parler***, etc. Le deuxième, appelé *pied-piping*, est le fait que le mot *qu-* (pronom relatif, interrogatif, etc.) qui apparaît dans le groupe extrait n'en est pas nécessairement la tête syntaxique et que sa distance avec la tête de ce groupe est également potentiellement illimitée : *la voiture **avec laquelle** j'ai eu un accident, l'ami de Jean **avec la voiture de qui** j'ai eu un accident, ?la personne **avec la voiture de l'ami de qui** j'ai eu un accident,* etc. Ces deux types de dépendances non bornées sont généralement traités suivant des modalités différentes (voir Section 4.4). Dans plusieurs de mes travaux, à commencer par [C2], j'ai montré que ces deux phénomènes pouvaient être traités par le même concept, le *nucléus*, et qu'en plus une série d'autres phénomènes pouvaient également être traités en termes de nucléus.

L'idée du nucléus est la suivante : certaines chaînes de mots se comportent du point de vue de certains phénomènes comme un seul mot. Nous considérons deux types de nucléus : les *nucléus verbaux*, qui sont des chaînes de verbes (*vouloir parler, commencer à parler, vouloir que … parle*, etc.) ou de tournures verbales (*avoir besoin de parler, être heureux de parler*, etc.) liés par des prépositions et des conjonctions de subordination, et les *nucléus nominaux*, qui sont des chaînes de noms reliés par des prépositions (*qui, avec qui, avec la voiture de qui*, etc.) (pour des définitions plus précises, voir [A2], [A3] et [A5]). Voyons maintenant comment un phénomène d'extraction, par exemple la relativisation, peut être traité en termes de nucléus. Dans la plupart des modélisations, on considère que n'importe quelle position d'une proposition peut être relativisée, puis on exclue ensuite les cas où l'élément relativisé doit franchir certaines frontières de syntagmes (cela est vrai pour les descriptions en termes de mouvement comme pour les descriptions en termes de traits SLASH à la G/HPSG) : par exemple, la frontière d'un modifieur ne peut être franchie (**la personne à qui Pierre regarde la télé sans parler* vs. *Pierre regarde la télé sans parler à Jean*). Mon approche est à l'opposé : elle considère que seul un dépendant direct du verbe principal d'une proposition peut être relativisé. Sauf que par verbe principal, on entend en fait nucléus verbal principal, c'est-à-dire n'importe quelle chaîne capable de faire office de verbe et que par dépendant du verbe, on entend nucléus nominal, c'est-à-dire n'importe quelle chaîne capable de se comporter comme un nom seul. Prenons un exemple traité dans [A3] :

(7) la dame sur le mari de laquelle je pense que tu peux compter

Dans la relative (7), le nucléus verbal principal est *pense que … peux compter* et le groupe extrait est le nucléus nominal *sur le mari de laquelle*. Cet exemple est illustré par la Figure 20 où le nucléus verbal (νV) et le nucléus nominal (νN) sont représentés par des bulles.

Conformément à la discussion de la Section 4.1, le pronom relatif *laquelle* est associé à deux semi-nœuds, l'un occupant la tête de la relative et assurant la translation du verbe principal de la relative (*pense*) et l'autre occupant une position dans la relative (comme dépendant de *mari de*).

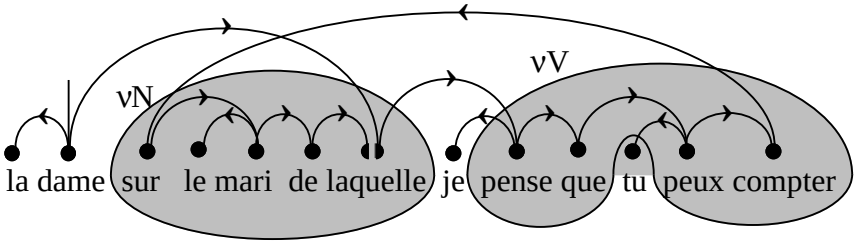


Figure 20 Arbre de dépendance de (7) avec indication des nucléus

Les nucléus ne sont pas une notion spécifique à l'extraction. Ceux-ci interviennent dans d'autres phénomènes syntaxiques, comme la coordination elliptique ou la négation. La *coordination elliptique* est le fait que, lorsque deux propositions ayant le même verbe principal sont coordonnées, le verbe principal de la deuxième puisse être effacé : *Pierre **parle** à Marie et Anne à Jacques*. Ici encore, on peut remplacer verbe principal par nucléus verbal principal : *Pierre **a l'intention de parle**r à Marie et Anne à Jacques, Pierre **veut qu'on appelle** la police et Anne les pompiers*. Dans la négation par un pronom négatif tel que *personne*, le verbe dont dépend le pronom reçoit le clitique *ne* : *Pierre ne **parle** à personne*. Ici encore, on peut remplacer verbe par nucléus verbal : *Pierre n'**a l'intention de parler** à personne, ?Pierre ne **supporte que je parle** à personne*[58]. Et l'on peut remplacer nom par nucléus nominal : *Pierre n'a parlé **aux parents de personne***.

D'autres phénomènes encore font intervenir des nucléus comme la comparaison ou l'accord. De plus, la notion de nucléus (c'est-à-dire le fait qu'un groupe de mots se comporte comme un seul mot vis-à-vis de certaines constructions) est probablement universelle, bien que chaque langue ait ses propres nucléus. Toutefois, une étude consacrée au concept de nucléus rassemblant des données dans un grand nombre de langues reste à faire.

Nous allons maintenant nous intéresser à la modélisation de l'extraction et la formalisation du nucléus qui la sous-tend.

4.3 Les arbres à bulles et la modélisation de l'extraction en GUST

Dans la Figure 20, nous avons représenté la structure de dépendance d'une relative en indiquant les nucléus. Dans cette représentation, les nucléus sont définis à partir de la structure de dépendance, laquelle prend comme éléments de base les mots et eux seuls. Dans [C2], j'ai proposé une structure mathématique nouvelle, *l'arbre à bulle*, permettant de traiter

[58] Cette phrase doit s'entendre avec le sens de 'Pierre ne supporte pas que je parle à quelqu'un' et ne peut être confondue avec la phrase *Pierre supporte que je ne parle à personne*, où la portée de la négation se limite à *parle*.

les nucléus comme des concepts primitifs : le nucléus n'est plus décrit à partir d'un formalisme donné, il est un élément à part entière de la structure, à savoir une bulle. On trouve des bulles dans les stemmas de Tesnière (pour la représentation de ce qu'il appelle le nucléus translatif). La seule formalisation de structures du type des arbres à bulles que je connaisse se trouve dans Gladkij 1968 (un mathématicien russe qui a par ailleurs collaboré avec I. Mel'čuk). Les arbres à bulles sont également utilisables pour la description des coordinations : les conjoints d'une coordination sont regroupés en une bulle et peuvent ainsi occuper un même nœud dans la structure et partager des dépendants. Les différentes combinaisons entre extraction et coordination sont décrites dans [C2] et les contraintes qui s'y rattachent sont exprimées par des propriétés géométriques simples de l'arbre à bulles (à la différence des grammaires syntagmatiques où ces contraintes font l'objet de propriétés *ad hoc*). Cet article contient les germes d'une idée qui me semble importante : une bonne représentation syntaxique permet de modéliser certaines propriétés de la langue par des propriétés *géométriques* de la structure de cette représentation. Cette idée est déjà présente dans les travaux des génératives, mais ceux-ci n'en tirent, à mon avis, qu'un parti limité, car les représentations syntaxiques qu'ils considèrent sont basées sur des structures mathématiques trop simples, des arbres (syntagmatiques), ce qui les oblige à considérer des arbres très complexes en termes de nombre de nœuds, où, de plus, une grande partie de l'information est cachée dans l'étiquetage des nœuds (par exemple la relation de tête avec les étiquettes X, X', X'' ou les mouvements avec les indices sur les traces) ou bien se trouvent dans des structures annexes sans géométrie explicite (théta-structure, théorie du cas, …).

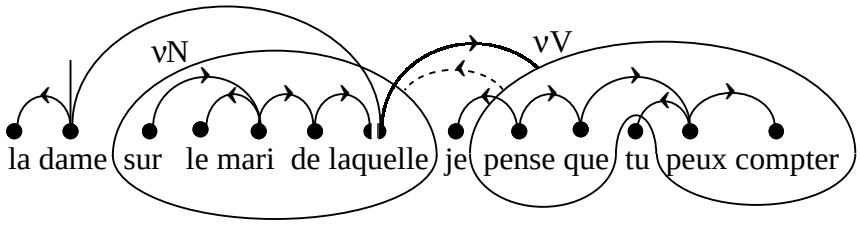


Figure 21 Un arbre à bulles pour (7)

Dans la Figure 21, nous proposons un arbre à bulles pour la relative (7)[59]. Ce n'est plus le verbe principal de la relative qui dépend du mot *qu-*, mais le nucléus verbal. De même, la dépendance entre *compter* et *sur* de la Figure 20 devient ici une dépendance entre le nucléus verbal et le nucléus nominal. Cette dépendance est indiquée en pointillé, car elle ne joue pas de rôle dans la linéarisation. Le placement du groupe extrait est entièrement assuré par la partie translative du mot *qu-* : comme tout translatif analytique du français, celui-ci doit être placé avant l'élément qu'il translate. Par le fait que le mot *qu-* forme bien un seul mot en

[59] Cet arbre à bulle est extrait de l'article [C13] qui n'est pas dans le recueil d'article joint à ce document de synthèse. L'article [C13] reprend en grande partie [A3], avec comme différence essentielle le fait qu'une double position est attribuée au mot *qu-* comme préconisé dans la dernière section de [A3].

surface, l'ensemble du groupe extrait se trouve alors dans la position imposée à la partie translative du mot *qu-*.

En autorisant les bulles représentant les nucléus à occuper des nœuds dans la représentation, l'extraction perd son caractère "non bornée", c'est-à-dire que la description de la relativisation peut être fait par des règles locales : une position peut être relativisée si elle appartient à un nucléus nominal qui dépend d'un nucléus verbal principal, c'est-à-dire un nucléus contenant le verbe principal de la relative. En un sens, cela revient à dire que seuls les dépendants du verbe principal peuvent être extraits, sachant que verbe vaut aussi pour nucléus verbal et que dépendant d'un verbe vaut pour nucléus nominal.

Avec la considération des nucléus, la description de la relativisation devient élémentaire : la difficulté, s'il en reste, est déplacée sur la description des nucléus verbaux et nominaux.

Il est possible d'étendre une grammaire de dépendance du type de GUST pour qu'elle manipule des arbres à bulles. L'analyse de l'extraction par des arbres à bulles a été intégrée à GUST pour donner une *grammaire à bulles* ([A3], [C13]). L'introduction de bulles dans les structures élémentaires et le fait que certains éléments peuvent dépendre d'une bulle nécessitent d'avoir des mécanismes permettant d'attribuer une dépendance à une bulle lors de la combinaison de deux structures élémentaires. Lorsqu'un nœud B se combine avec un nœud A appartenant à un nucléus, trois cas sont envisageables : le nœud B se retrouve à l'extérieur du nucléus (c'est la règle par défaut), le nœud B est placé dans le nucléus (la dépendance entre A et B est dite *intranucléaire*) ou le nœud B "monte" sur le nucléus (la dépendance entre B et le nucléus est dite *extranucléaire*). Les dépendances potentiellement intra ou extranucléaires sont répertoriées et ceci pour chaque type de nucléus : c'est de cette façon qu'est contrôlée la bonne formation des nucléus. Dans [A3], où est présentée une version lexicalisée de GUST, les liens potentiellement intra et extranucléaires sont indiqués directement dans les structures élémentaires associées aux lexèmes (voir les étiquettes iνV, eνV et iνN dans la Figure 22). Les opérations de combinaison avec nucléus peuvent être mis en parallèle avec les opérations utilisées en HPSG : le fait qu'une dépendance devienne extranucléaire (et soit attribuée au nucléus verbal) correspond à l'extraction proprement dite et est équivalente au fait de mettre un élément dans le trait SLASH, tandis que le fait qu'une dépendance devienne intranucléaire correspond à la remontée de l'élément extrait et est équivalent à la percolation du trait SLASH. Par exemple, dans la Figure 22, la dépendance entre COMPTER et SUR devient extranucléaire (eνV), ce qui reviendrait en HPSG à mettre la requête SUR dans le trait SLASH de COMPTER ; la dépendance entre POUVOIR et COMPTER devient intranucléaire (iνV), ce qui reviendrait en HPSG à remonter le contenu du trait SLASH de COMPTER sur POUVOIR. La différence essentielle avec HPSG est que, en GUST, rien ne remonte par POUVOIR et que la structure associée à POUVOIR reste la même que s'il n'y avait pas d'extraction (la seule différence pour POUVOIR est qu'il se trouve dans un nucléus verbal). Ce qui est propagé, ce n'est pas le contenu du groupe extrait, mais l'appartenance au nucléus verbal, c'est-à-dire le pouvoir de gouverner le groupe extrait. De plus, la version grammaire à bulles de GUST reste associative, c'est-à-dire que l'ordre dans

lequel les structures élémentaires de GUST se combinent est libre, alors qu'en HPSG, on doit suivre une stratégie de bas en haut (*bottom-up*) et SUR doit être dans le trait SLASH de COMPTER avant que POUVOIR se combine avec COMPTER (voir [A7]).

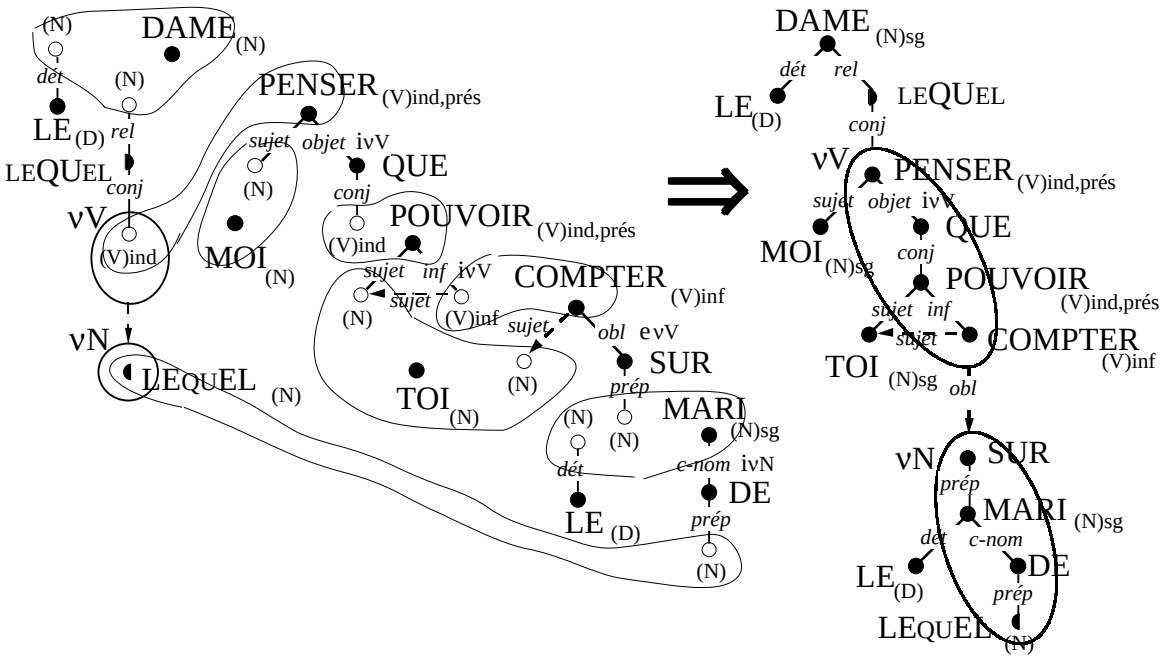


Figure 22 Dérivation de (7)

La Figure 22 est extraite de [C13]. Dans [C13], comme dans [A3], l'architecture de GUST n'est pas encore bien dégagée et les structures élémentaires ne sont pas considérées comme des signes effectuant l'interface entre deux niveaux de représentation. En fait les structures élémentaires qui apparaissent dans la partie gauche de la Figure 22 sont des signifiants de signe profond et si on y ajoute la partie signifiée — les relations prédicat-argument — la combinaison de ces structures réalise l'interface sémantique-syntaxe.

Nous allons nous concentrer sur le signe LEQUEL, repris dans la Figure 24. Comme on le voit dans structure associée à LEQUEL, nous attachons à ce signe lexical l'ensemble de la description propre à la relative : c'est dans la structure de LEQUEL qu'il est dit qu'une relative est une proposition (la tête est un verbe à l'indicatif) et qu'une relative modifie un nom et c'est aussi dans cette structure qu'est contrôlée l'extraction. En effet, la structure de LEQUEL introduit les nucléus verbal et nominal et assure que la position relativisée appartient à un nucléus nominal dépendant d'un nucléus verbal principal. Il pourrait être judicieux, dans une version moins lexicaliste de la grammaire, d'isoler la partie de la structure propre à la construction des relatives et en faire un signe indépendant comme on l'a fait pour Adj À Vinf (Figure 13).

Voyons maintenant la sémantique du signe profond LEQUEL. Les pronoms relatifs n'introduisent pas de sémantème propre. Ils ont un rôle anaphorique et reprennent la sémantique de l'antécédent de la relative. Dans la partie gauche de la Figure 23, nous donnons la représentation sémantique simplifiée de (7) : on y voit les relations prédicat-argument entre

les signifiés des différents signes profonds de (7) à l'exception du pronom LEQUEL qui n'introduit pas de sémantème. Conformément à [A2], le contenu de la relative est isolé et le sémantème 'penser' qui sera réalisé comme la tête de la relative est marqué comme nœud dominant. La sémantique des relatives est ici traitée de façon extrêmement grossière et nous considérons seulement la question de l'organisation entre les éléments de la structure sémantique et pas leur nature propre.

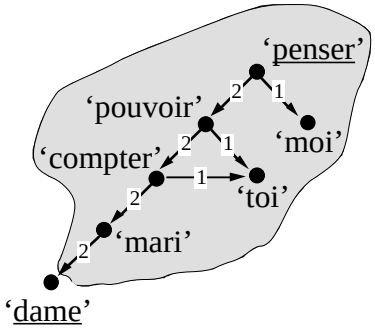


Figure 23 Représentation sémantique de (7)

Nous donnons Figure 24 (côté gauche) la structure élémentaire associée au signe profond LEQUEL ; la question du calcul du sous-graphe correspondant au contenu de la relative n'est pas formalisé (nous avons juste indiqué en toutes lettre que le verbe principal de la relative doit être le nœud dominant du sous-graphe correspondant au contenu de la relative). La Figure 24 (côté droit) donne la structure de dérivation correspondant à la dérivation de (7) proposée dans la Figure 22. On y voit que LEQUEL se combine avec trois structures élémentaires : l'antécédent de la relative DAME, le verbe principal de la relative PENSER et son gouverneur dans la relative MARI. Non seulement LEQUEL crée un cycle dans la structure de dérivation, mais il crée également un cycle dans la structure hiérarchique sous-jacente à la structure de dérivation, puisque LEQUEL à la fois domine et est dominé par PENSER en raison de son double rôle syntaxique.

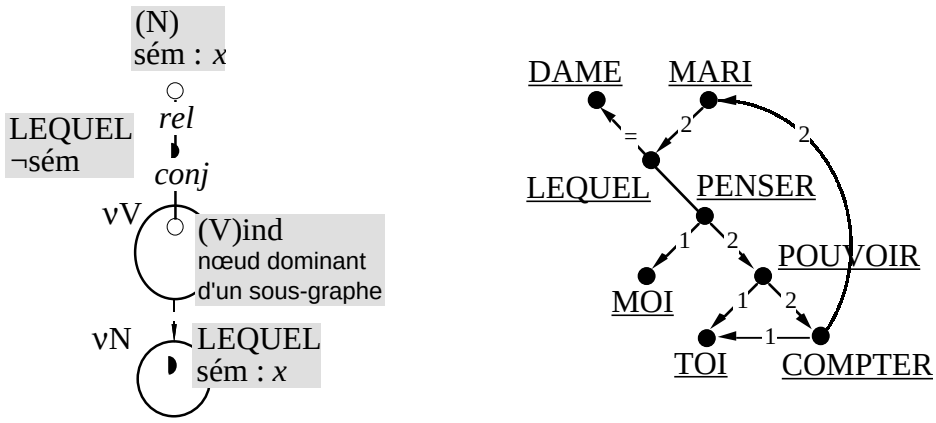


Figure 24 Structure élémentaire associée à LEQUEL (gauche)
et structure de dérivation correspondant à la Figure 22 (droite)

Terminons par l'interface syntaxe-morphologie, c'est-à-dire la question de l'ordre des mots. L'ordre des mots dans une phrase à extraction est considéré comme complexe en raison de la non-projectivité de la structure de dépendance, laquelle résulte du fait que, en général, le groupe extrait n'est pas dans la projection de son gouverneur syntaxique. Comme on le voit dans la Figure 20, la dépendance entre *compter* et *sur* (la tête du groupe extrait) recouvre plusieurs ancêtres de *compter* et coupe la dépendance entre *dame* (l'antécédent de la relative) et la tête de la relative. Dans l'analyse en termes d'arbre à bulles donnée Figure 21, la dépendance entre *compter* et *sur* devient une dépendance entre le nucléus verbal et le nucléus nominal et la structure est maintenant projective : il n'y a plus de coupure entre deux dépendances. Du fait que la structure est projective, le placement peut être calculé en utilisant uniquement des règles d'ordre locales, c'est-à-dire des règles qui mettent en jeu un nœud et ses fils. De plus, la linéarisation ne nécessite pas de règle spécifique à l'extraction. En particulier, le placement du groupe extrait au début de la relative découle directement de la présence du mot *qu-* dans le groupe extrait et de son rôle translatif, car, comme tous les translatifs analytiques du français (prépositions, conjonctions, etc.), le mot *qu-* doit être avant l'élément qu'il translate, c'est-à-dire le verbe principal de la relative. Comme on le voit, le fait d'attribuer deux positions syntaxiques au mot *qu-* et en particulier le fait de faire apparaître sa position de translatif simplifie l'interface syntaxe-morphologie.

Nous allons maintenant voir comment la modélisation des extractions en GUST peut être transférée dans d'autres formalismes.

4.4 Autres formalisations de l'extraction

La modélisation de l'extraction que nous venons de présenter, basée sur une double position syntaxique du mot *qu-* et sur les concepts de nucléus, a été implantée dans d'autres cadres, notamment TAG et HPSG, servant à chaque fois de point de comparaison entre les théories et/ou les formalismes.

Concernant les TAG, nous avons proposé dans [C9] une nouvelle analyse de l'extraction basée sur l'analyse de GUST. Avant de présenter cette analyse, rappelons le principe de l'analyse classique de Kroch & Joshi 1986. Dans leur analyse, la structure du verbe qui gouverne la position extraite s'adjoint sur l'antécédent (et contrôle donc la distribution de la relative). Le mot *qu-* est ensuite substitué dans cette structure. Pour assurer qu'il y ait bien un mot *qu-* dans la relative, la position à substituer possède un trait wh:+ qui bloque un autre type de mot. Pour permettre le pied-piping, un nom peut être autorisé à occuper une position wh:+ à condition de faire suivre à l'une de ses propres positions le trait wh:+. Par exemple, dans l'analyse de (7), la structure élémentaire de *compter sur* va s'adjoindre sur l'antécédent *dame*. La position à substituer sous *sur* (dans l'arbre de *compter sur*) est marquée wh:+. La structure associée à *le mari de* va venir s'y substituer et proposera à son tour une position à substituer marquée wh:+ sous *de*, qui sera finalement occupée par le pronom *laquelle*. Par ailleurs, les

structures associées à *peux* et *je pense que* vont s'adjoindre l'une après l'autre sur (*tu*)[60] *compter sur*. L'analyse de Kroch & Joshi 1986 pose au moins trois problèmes :

- La structure élémentaire associée au verbe qui gouverne la position extraite (ici *compter sur*) s'adjoint sur l'antécédent, alors que celui-ci n'est pas l'un de ses arguments sémantiques. Ceci est contraire aux principes linguistiques qui gouvernent les grammaires TAG et donne un caractère *ad hoc* à l'analyse de l'extraction en TAG. Il en découle que l'arbre de dérivation ne peut plus être interprété comme une représentation sémantique (cf. [C6]), car l'arc associé à l'adjonction sus-citée ne s'interprète pas comme une relation prédicat-argument.
- La structure élémentaire associée au verbe qui gouverne la position extraite est spécifique à la relativisation. Une autre structure élémentaire pour le même verbe est nécessaire pour l'interrogation indirecte ou pour la topicalisation. Cette analyse de l'extraction contribue donc fortement à l'explosion combinatoire du nombre de structures élémentaires associées à un même verbe (cf. [C10]).
- Les deux types de dépendances non bornées mises en jeu dans l'extraction, c'est-à-dire l'extraction proprement dite et le pied-piping, sont traités par des moyens formels très différents. En particulier, le pied-piping est traité par la percolation d'un trait (le trait wh:+), ce qui est contraire à l'esprit de TAG où la géométrie des opérations formelles se reflète normalement dans l'arbre de dérivation.

La nouvelle analyse proposée dans [C9] résout partiellement ces trois points. Elle repose sur le double rôle pronominal et translatif du mot *qu-* et c'est donc le mot *qu-* qui contrôle la distribution de la relative et s'adjoint sur l'antécédent de la relative. Ensuite tous les éléments de la chaîne de dépendance de la position occupée par le mot *qu-* dans la relative jusqu'au verbe principal de la relative vont s'adjoindre un à un à partir du mot *qu-*. Les avantages de cette analyse sont immédiats :

- Les principes linguistiques sous-jacents à la modélisation des langues par des TAG sont respectés et la structure de dérivation donne les dépendances sémantiques attendues, ce qui simplifie grandement l'interface sémantique-syntaxe.
- Le pouvoir de s'adjoindre sur l'antécédent ayant été transféré sur le pronom relatif, les structures élémentaires associées aux verbes s'en trouvent simplifiée et, qui plus est, la même structure peut servir pour les différents types d'extraction (relativisation interrogation, etc.) réduisant le nombre de structures élémentaires associées à un verbe.

[60] Il est d'usage en TAG de privilégier les dépendances sémantiques sur les dépendances syntaxiques (voir [C6]). Ainsi dans une construction comme *tu peux compter* [*sur le mari de cette dame*], la structure de *tu* sera substituée dans celle de *compter*, dont il est l'argument sémantique, plutôt que dans celle de *peux*, dont il est seulement sujet syntaxique. Un tel choix n'a pas lieu en GUST où la structure de TOI se combine simultanément avec celles de POUVOIR et COMPTER.

- Les deux types de dépendances non bornées mises en jeu dans l'extraction sont traitées de la même façon et l'usage de la percolation de trait est évité.
- La couverture de la grammaire est augmentée (par exemple, certaines relatives en *dont* du français peuvent maintenant être traitées).

Il reste des cas problématiques, notamment ceux où le pronom relatif ou le groupe extrait a un rôle de modifieur comme dans *le lit où je dors*, *le fauteuil dans lequel je suis assis*. Le problème vient du fait que le pronom *où* devrait à la fois s'adjoindre sur l'antécédent de la relative *lit* et sur le verbe *dors* qu'il modifie. Or le formalisme de TAG ne permet pas la double adjonction.

Avec Marie-Hélène Candito, nous avons proposé, dans [C7], un formalisme à la TAG où les adjonctions multiples sont possibles. Les structures de dérivation dans ce formalisme ne sont plus des arbres (une structure élémentaire qui s'adjoint sur deux structures aura deux gouverneurs), mais des graphes hiérarchisés, ce qui nous a conduit a appelé ce formalisme *GAG* pour *Graph-driven Adjunction Grammar*. Les structures de dérivation de ce formalisme sont très proche de celle de GUST. Je ne présenterais pas davantage ce formalisme, ni l'analyse que nous proposons pour l'extraction, car ce formalisme pose les mêmes problèmes majeurs que TAG, en particulier le fait de ne pas être suffisamment articulé et de réaliser par un seul module l'interface sémantique-morphologie.

Passons maintenant à l'analyse de l'extraction en HPSG que j'ai proposé dans [A7]. Dans la présentation que j'ai faite de la modélisation de l'extraction en GUST (Section 4.3), j'ai déjà évoqué le lien entre les opérations utilisées en GUST pour la combinaison des arbres à bulles et les opérations utilisées en HPSG pour le traitement des dépendances non bornées, à savoir la percolation de traits. Ce lien a été mis en évidence pour la première fois dans [C1], mais la présentation concerne davantage les grammaires de dépendance que HPSG. Dans [A7], je m'intéresse avant tout à HPSG et je propose une nouvelle modélisation de l'extraction dans cette théorie. Cette modélisation se distingue des analyses traditionnelles d'HPSG (Pollard & Sag 1994, Bouma *et al.* 2001) par le fait que toute l'information linguistiques est introduite dans les structures associées aux mots et que les schémas syntagmatiques ne font que gérer les combinaison de trait et la percolation. Comme en GUST, un double rôle est attribué au mot *qu-* et c'est lui qui contrôle la distribution de la proposition et, par exemple, le rattachement d'une proposition relative à son antécédent. Le schéma *head-filler-phrase* proposé s'interprète uniquement en termes de combinaisons de mots, à la différence du schéma traditionnel pour lequel des informations linguistiques cruciales sont directement introduites au niveau du constituant. Ce nouveau schéma est plus complexe que les schémas du type *head-comp-phrase* qui assurent la combinaison d'une tête avec un de ses dépendants, car il met en jeu une double combinaison due au double rôle du mot *qu-* : tandis que le groupe extrait (*filler*) remplit la position extraite (*gap*), le mot *qu-* se combine avec le verbe principal de la proposition et assure la bonne distribution de cette proposition. Le fait que les schémas phrastiques ne fassent que manipuler les informations des constituants filles et n'introduisent pas eux-mêmes de l'information linguistique fait de cette version d'HPSG une grammaire

complètement lexicalisée et l'a rend équivalent à une grammaire de dépendance. Par ailleurs, comme je l'ai mentionné dans la Section 1.4, le formalisme d'HPSG n'est pas associatif comme celui de GUST. L'analyse HPSG des extractions avec trait SLASH correspond à l'analyse de GUST où l'arbre syntaxique est construit de bas en haut. D'autres ordres dans la combinaison des signes profonds de GUST peuvent être proposés : une procédure de combinaison davantage guidée par l'ordre linéaire et donnant une grammaire HPSG sans SLASH est proposée dans la dernière partie de [A7] (et esquissée dans la Section 1.4).

En conclusion, l'analyse de l'extraction avec nucléus et double rôle du mot *qu-* a pu être adaptée à différents formalismes, en donnant à chaque fois une modélisation plus élégante que la modélisation originale. La modélisation la plus satisfaisante reste néanmoins celle de GUST qui repose sur les arbres à bulles et où les nucléus sont ainsi traités comme des concepts primitifs.

5 Ordre des mots et constituants de surface

Dans la Section 3.3, nous avons proposé un premier traitement de l'ordre des mots particulièrement simple, mais insuffisant pour traiter les constructions non projectives et inadapté pour traiter les langues dites à ordre libre, où l'ordre des mots dépend davantage de la structure communicative que des relations syntaxiques entre les mots.

Avec Kim Gerdes, nous nous sommes intéressés à la syntaxe de l'allemand et un peu du néerlandais. Les langues germaniques ont la particularité d'être des langues à ordre libre tout en obéissant à des schémas syntaxiques très contraignants. La syntaxe de ces langues présente un grand nombre de phénomènes considérés comme complexes et qui ont fait l'objet d'une littérature abondante : scrambling, VP fronting, extraposition, Oberfeldumstellung (auxiliary flip), cross serial dependencies, etc. Dans [C16], nous proposons une description de l'ensemble de ces phénomènes basée sur une hiérarchie de domaines divisés en champs selon une adaptation du *modèle topologique* de Drach 1937 et Bech 1955 (voir également Kathol 1995 pour une implantation du modèle topologique en HPSG). Selon le modèle topologique, la phrase déclarative allemande possède cinq champs appelés *Vorfeld* (champ de début), *parenthèse gauche*, *Mittelfeld* (champ du milieu), *parenthèse droite* et *Nachfeld* (champ de fin). Le verbe principal occupe toujours la parenthèse gauche, c'est-à-dire la deuxième position de la phrase, car le Vorfeld accueille un et un seul constituant. Les dépendants non verbaux se placent dans l'un des trois champs majeurs : Vorfeld, Mittelfed ou Nachfeld (avec quelques restrictions pour le Nachfeld). Un dépendant verbal (participe passé, infinitif nu ou infinitif en *zu*) a lui deux possibilités nettement distinctes :

La première possibilité pour un dépendant verbal est de rejoindre la parenthèse droite du domaine de son gouverneur. Dans ce cas, ses dépendants appartiennent au même domaine que ceux de son gouverneur et peuvent se "mélanger" (angl. *to scramble*) avec eux librement, comme en (8). Un dépendant verbal d'un verbe de la parenthèse droite peut rejoindre son

gouverneur dans la parenthèse droite, formant avec lui un amas verbal (angl. *verb cluster*)[61]. Aucun élément non verbal ne peut occuper la parenthèse droite, à l'exception de certaines particules verbales qui se comportent exactement comme des verbes et de certains éléments prédicatifs (nom ou adjectif).

(8)

Vorfeld	*(*	*Mittelfeld*	*)*	*Nachfeld*
Niemand	hat	den Roman diesem Mann	zu lesen versprochen	
Personne (nom.)	a	le roman (acc.) à cet homme (dat.)	de lire promis	

'Personne n'a promis de lire le roman à cet homme'

- La deuxième possibilité pour un dépendant verbal est de se placer dans les domaines majeurs. Dans ce cas, le verbe ouvre un domaine enchâssé formé d'un Mittelfeld, une parenthèse droite (que le verbe occupe) et un Nachfeld. Les dépendants de ce verbe peuvent rejoindre ce domaine (comme en (9)) ou s'émanciper et rejoindre un domaine supérieur (comme en (10)).

(9)

Vorfeld			*(*	*Mittelfeld*	*)*	*Nachfeld*
MF	*)*	*NF*				
Den Roman	zu lesen		hat	diesem Mann niemand	versprochen.	
le roman (acc.)	de lire		a	à cet homme (dat.) personne (nom.)	promis	

(10)

Vorfeld			*(*	*Mittelfeld*	*)*	*Nachfeld*
MF	*)*	*NF*				
	Zu lesen		hat	diesem Mann den Roman niemand	versprochen.	
	de lire		a	à cet homme(dat.) le roman(acc.) personne(nom.)	promis	

Notre analyse repose sur l'idée que selon le champ dans lequel il est placé, un verbe va ou non ouvrir un nouveau domaine, ce qui veut dire que la nature même du constituant ouvert par un verbe ne dépend pas seulement de sa position syntaxique (c'est-à-dire de son gouverneur syntaxique), mais aussi de sa position linéaire (ou, plus exactement, de sa position topologique). Cette idée nouvelle et néanmoins fort simple s'oppose complètement aux

[61] En général, dans la parenthèse droite, le dépendant doit se placer à gauche de son gouverneur. Les auxiliaires du passé et du futur, HABEN 'avoir' et WERDEN 'devenir', permettent à leur dépendant de se placer à droite (le phénomène est limité à une classe fermée de verbes contenant les modaux et quelques autres verbes). Le phénomène est connu sous le nom d'*auxiliary flip*. Le phénomène possède lui-même deux variantes suivant que le dépendant V_3 du deuxième verbe se place à la gauche immédiate de son gouverneur (Oberfeldumstellung ; $V_1V_3V_2$) ou à la gauche de l'amas complet (Zwischenstellung ; $V_3V_1V_2$).

approches basées sur la syntaxe X-barre (Jackendoff 1977). En particulier, dans notre approche, à l'inverse des grammaires syntagmatiques traditionnelles, deux phrases ayant la même structure syntaxique, c'est-à-dire le même arbre de dépendance, pourront avoir des structures de constituants très différentes.

Nous avons formalisé cette analyse dans une grammaire de dépendance d'un type nouveau, alternant des règles d'assignation de place dans des champs et des règles de création de constituants en fonction des champs ([C17]). Pour être plus précis, notre grammaire associe un arbre de dépendance syntaxique à une structure topologique. Les symboles manipulés par la grammaire correspondent aux catégories syntaxiques, aux relations syntaxiques, aux constituants topologiques (domaines, amas verbal, groupe nominal, …) et aux champs. Nous avons quatre types de règles :

- des *règles de description de constituant*, qui à un nom de constituant associent une liste de champs ; par exemple, au domaine principal correspond une liste de cinq champs : [Vorfeld, parenthèse gauche, Mittelfeld , parenthèse droite, Nachfeld] ; à l'amas verbal correspond une liste de trois champs : [Oberfeld, champ de tête, Unterfeld] :
- des *règles de description de champ*, qui indiquent si un champ peut ou non rester vide et s'il peut contenir un seul constituant ou un nombre quelconque ; par exemple le Vorfeld doit contenir un et un seul constituant, tandis que les Mittelfeld et Nachfeld peuvent en contenir un nombre quelconque ;
- des *règles de correspondance* (ou *syntagmes*), qui indiquent dans quel champ peut aller un élément d'un catégorie donnée en fonction de la catégorie de son gouverneur et de sa fonction syntaxique ; ces règles gèrent également l'émancipation en indiquant quelles frontières de constituants peuvent séparer le gouverneur de son dépendant ; par exemple, un dépendant d'un verbe peut aller dans un champ majeur en traversant éventuellement un frontière de domaine enchâssé ;
- des *règles de création de constituant*, qui indiquent quel type de constituant crée un mot d'une catégorie donnée quand on le place dans un champ donné et dans quel champ de ce constituant il se place ; par exemple, un verbe placé dans un champ majeur crée un domaine enchâssé et occupe la parenthèse droite, tandis qu'un verbe placé dans une parenthèse droite crée un amas verbal et occupe le champ de tête[62].

Nous avons développé dans ce formalisme une grammaire pour la syntaxe du verbe en allemand, ainsi qu'un traitement du pied-piping et notamment du pied-piping verbal, qui n'existe pas en français et en anglais (voir exemple en [C17]).

Nous avons également réalisé, en collaboration avec Lionel Clément qui développe un analyseur LFG, une implantation de cette analyse dans le formalisme LFG avec quelques

[62] La deuxième règle sera déclenchée à la suite de la première, puisqu'un verbe qui crée un domaine enchâssé, crée ensuite un amas verbal dans la parenthèse droite de ce domaine. La deuxième règle peut aussi être utilisée seule quand un verbe est envoyé dans la parenthèse droite du domaine de son gouverneur.

aménagements théoriques de LFG, notamment en introduisant la notion de champ à coté de celle usuelle de constituant ([C19]).

Kim Gerdes a réalisé une implantation de notre grammaire de l'allemand avec un synthétiseur qui prend en entrée un arbre de dépendance et fournit tous les ordres possibles et toutes les structures de constituants possibles. Nous avons ajouté sur nos règles un système de pénalités pour classer les structures par ordre de complexité (et donc d'acceptabilité) dont les résultats sont très satisfaisants (http://talana.linguist.jussieu.fr/~kim/deplin). Le principe est de pénaliser certaines opérations comme la création d'un domaine enchâssé dans le Mittelfed ou une émancipation. Nous prenons ainsi le contre-pied des analyses transformationnelles où la structure de base est une structure où chaque verbe crée un VP enchâssé et où les phrases à un seul domaine sont obtenues par un grand nombre de mouvements remontant tous les compléments dans le VP principal.

En plus de permettre le calcul des différents ordres des mots possibles, notre grammaire de l'allemand construit une hiérarchie de domaines. Cette structure de constituants, que nous appelons la *structure topologique*, diffère nettement de la structure de constituants des grammaires syntagmatiques basée sur la Syntaxe X-barre. En fait, la structure syntagmatique usuelle cherche à encoder le même type d'information qu'un arbre de dépendance, c'est-à-dire les relations de dépendance (sous-catégorisation, modification et coordination) ; elle en diffère par le fait qu'elle intègre l'ordre des mots, ce qui ne va pas sans quelques difficultés. Notre structure de constituants formée des domaines se place à un niveau plus surfacique où l'ordre des mots est considéré en tant que tel. Cette structure de constituants rend compte des regroupements de mots qui sont certes imposés par la structure de dépendance, mais également et surtout par la structure communicative (la partition thème-rhème, les mises en contraste, en arrière-plan, etc.).

Dans [C18], nous nous sommes attachés, toujours avec Kim Gerdes, à motiver davantage la structure topologique. Le fait de pouvoir décrire l'ensemble des phénomènes d'ordre des mots liés à la syntaxe des verbes avec un petit nombre de règles simples comme en [C16] pourrait constituer en soi une validation des approches basées sur le modèle topologique, mais cela n'explique pas quelle est la nature exacte des constituants topologiques. L'une des particularités de l'analyse topologique est que certaines phrases sont topologiquement ambigues. Par exemple, la phrase (11) possède trois analyses topologiques : dans la première, il n'y a qu'un domaine et *zu lesen versprochen* forme un amas verbal dans la parenthèse droite de ce domaine ; dans la deuxième, *das Buch zu lesen* forme un domaine enchâssé dans le Mittelfeld du domaine principal et enfin dans la troisième (fortement pénalisée et donc très improbable), *zu lesen* forme un domaine enchâssé dont *das Buch* s'est émancipé.

(11)	*Niemand*	*hat*	*diesem Mann*	*das Buch*	*zu lesen*	*versprochen*
	Personne	a	cet homme	le livre	de lire	promis

Nous avons pu montrer que la phrase (11) possède deux prosodies naturelles correspondant aux deux premières structures topologiques sus-décrites : dans la première, l'amas verbal *zu*

lesen versprochen est nettement présent avec un seul accent sur la première syllabe accentuable (*zu lesen versprochen*), alors que dans la deuxième, le groupe *das Buch zu lesen* est légérement détaché et *versprochen* reçoit un accent sur la première syllabe de son radical (*versprochen*). Nous voyons donc que la structure topologique reflète de façon significative les groupements prosodiques et est par conséquent un niveau de structuration intermédiaire entre la structure communicative[63] et la structure prosodique.

Nous sommes maintenant en train de faire un travail comparatif sur la façon dont la structure communicative se réalise en russe, allemand et français. Il se trouve que dans toutes ces langues nous retrouvons à peu près les mêmes ordres des mots et les mêmes schémas prosodiques pour une structure communicative donnée. Le cas du français est assez amusant : l'ordre y est moins libre que dans les autres langues considérées, mais un certain nombre de constructions, notamment les dislocations gauche et droite, le clivage et le pseudo-clivage, permettent d'opérer les mêmes regroupements et d'obtenir en fait tous les ordres possibles. Illustrons notre propos par une construction à deux actants du russe :

(12) *Vanja* *prigotovil* *malinu*
Vania-NOM prépare framboises-ACC

En réponse à la question *Que prépare Vania ?* (c'est-à-dire avec *malinu* 'les framboises' comme rhème), les six ordres des mots de la phrase (12) sont possibles. Nous donnons en (13) ces six ordres avec la ou les prosodies[64] et avec la ou les constructions du français qui peuvent s'exprimer par cet ordre des mots en russe :

(13) **a.** *Vanja prigotovil malinu* --^ + - /,^ + écrit
Vania prépare les framboises
Ce que Vania prépare, c'est les framboises

b. *Prigotovil Vanja malinu* -/,^
Ce que prépare Vania, c'est les framboises

c. *Malinu prigotovil Vanj*a ^-- + écrit
C'est les framboises que prépare Vania

d. *Malinu Vanja prigotovil* ^--
C'est les framboises que Vania prépare

e. *Prigotovil malinu Vanja* /,^,-
Ce qu'il prépare, c'est les framboises, Vania

f. *Vanja malinu prigotovil* /,^-
Vania, c'est les framboises qu'il prépare

[63] La structure communicative n'est pas le seul déclencheur d'un domaine enchâssé. Par exemple, il est courant, en allemand, de rejeter un verbe dans le Nachfeld (où il ouvrira un domaine enchâssé) simplement pour alléger la phrase et faciliter la compréhension en n'éloignant pas exagérément la parenthèse droite de la parenthèse gauche.

[64] Nous indiquons les accents sur les trois groupes ; l'accent de focus porté par *malinu* est noté par ^ et l'accent de topic suivi d'un pause par /,. Les formes acceptables à l'écrit sont également mentionnées.

Comme on le voit, toutes les gloses du français en (13) sont des réponses acceptables à la question *Que prépare Vania ?* et elle possède à peu près la même prosodie que leur contrepartie russe avec le même ordre des trois segments *Vanja*/*Vania*, *malinu*/*les framboises*, *prigotovil*/*prépare*.

Inversement, pour un ordre donné, il est possible en français d'obtenir toutes sortes de regroupement communicatifs. Le cas de l'ordre canonique est le plus prolifique :

(14) **a.** *Pierre a mangé les framboises*

b. *Pierre les a mangées, les framboises*

c. *Pierre, il a mangé les framboises*

d. *Pierre, il les a mangées, les framboises*

e. *C'est Pierre qui a mangé les framboises*

f. *C'est Pierre qui les a mangées, les framboises*

g. *Ce que Pierre a mangé, c'est les framboises*

h. *Pierre, ce qu'il a mangé, c'est les framboises*

Le français a ceci de particulier que certains éléments semblent échapper au domaine de la rection. Dans une phrase comme *Moi, ce soir, c'est fini, les vacances*, ni *moi*, ni *les vacances* ne sont dans la valence du noyau central *c'est fini*. Pour reprendre les termes de Blanche-Benveniste 1975, 1997, Berrendonner 1991 ou Deulofeu 1989, on a d'une part une syntaxe nucléaire ou une *microsyntaxe*, qui obéit aux principes de la syntaxe dépendance et d'autre part une syntaxe de zones ou *macrosyntaxe*, avec des éléments disloqués sans lien de dépendance avec le noyau central[65] (voir également Van Valin & La Polla 1997 pour une distinction similaire entre éléments du "noyau" (angl. *core*) et éléments extérieurs). Notre modèle ne prend en compte que la microsyntaxe (c'est-à-dire la syntaxe de dépendance). Une modélisation de la macrosyntaxe et tout particulièrement de l'interface sémantique-syntaxe de ces phénomènes constitue un enjeu majeur.

Le travail sur l'ordre des mots en allemand avec Kim Gerdes s'est déroulé parallèlement à un début de travail avec Igor Mel'čuk sur l'ordre des mots en français. Contrairement au travail sur l'allemand où nous sommes partis de la phrase et de ses sous-constituants, les compléments verbaux, nous cherchons avec Igor Mel'čuk à caractériser les plus petits constituants du français. Nous sommes retombés sur une famille de constituants déjà populaires en linguistique computationnelle et en prosodie, les "chunks" ou syntagmes non récursifs (Abney 1991, Vergne 2000). Par exemple, une phrase du français comme *Il parle à la sœur de Jean* a pour découpage en chunks : (*il parle*) (*à la sœur*) (*de Jean*). Ce découpage reflète la façon dont les mots se regroupent en surface et les prosodies possibles pour cette phrase (en particulier les pauses possibles). Par contre, ce découpage ne rend pas compte de la

[65] On peut, comme le propose Mel'čuk 1988, faire dépendre ces éléments de la racine de l'arbre syntaxique, c'est-à-dire du verbe principal de la phrase, mais quoi qu'il en soit il s'agit de dépendances de nature différente des dépendances qui relient le verbe avec les éléments de sa valence.

structure de dépendance que cherche au contraire à saisir une structure syntagmatique standard telle que [(*il*) (*parle* (*à* (*la sœur* (*de* (*Jean*)))))]. Ces constituants sont, à notre connaissance, absents des travaux de linguistique théorique. Le tableau suivant indique à quelles conditions deux élément X et Y forment un constituant de type C. On voit qu'il existe entre les mots-formes et les groupes des types intermédiaires de syntagmes. La terminologie relative à ces syntagmes est déficiente : les points d'interrogation indique l'absence de terme approprié ou un terme contestable.

X = tête	Y	C	exemples	caractéristiques
racine	affixe	mot-forme	*re-faire, in-suffisant*	ordre rigide, pas de pauses, pas d'insertion
hôte	clitique	? mot prosodique	*il le lui donne*	ordre rigide, pas de pauses, insertion très contrôlée (clitiques seulement)
?	?	chunk	*le petit garçon*	ordre rigide, pas de pauses, pas de parenthétiques
?	?	quasi-chunk	*sauf ce livre* *une chemise de coton*	ordre rigide, pas de pauses hors parenthétiques
tête	dépendant	groupe	(*une robe*) (*de bonne qualité*) (*verte et jaune*)	ordre libre, pauses possibles

L'*ordre rigide* signifie que les différents éléments Y dépendants de X ont des places fixes les uns par rapport aux autres (ainsi dans le chunk nominal le déterminant précède les adjectifs et l'ordre de ceux-ci est lui-même très contraint : *un nouveau petit problème*, **un petit nouveau problème*). A noter également qu'en français la liaison est obligatoire dans le chunk et impossible hors des quasi-chunks. Le statut des éléments postposé au nom pose des problèmes intéressants. L'adjectif postposé va appartenir au chunk nominal, lorsqu'il est nu (*des livres intéressants*), mais pas lorsqu'il a un complément (*des livres intéressants à lire*). Par exemple, la liaison est possible dans le premier cas, mais pas dans le deuxième. Notons également que les groupes en *de* N où N est un nom nu forme probablement un chunk N *de* N avec leur gouverneur. En particulier, ce groupe doit se placer avant un adjectif postposé (*une chemise d'homme blanche* ; ?**une chemise blanche d'homme*). Et ce groupe est unique : par exemple, dans *une chemise pour homme en coton pour le sport*, l'ordre des trois compléments de *chemise* est libre et chacun d'eux peut donner un complément en *de* N (*une chemise d'homme* ; *une chemise de coton* ; *une chemise de sport*), mais le complément en *de* N est forcément unique et précède les autres (*une chemise d'homme en coton*, **une chemise d'homme de coton*, ?**une chemise pour homme de coton*).

A travers ces différents travaux sur l'ordre des mots, nous défendons une conviction profonde qui est que, contrairement à la tradition dominante des grammaires syntagmatiques, il est nécessaire de séparer clairement les dépendances des regroupements liés à l'ordre des mots

(même si les deux sont directement liés par les règles d'ordre des mots). Ces idées émergent également, bien que de façon encore confuse à mon avis, des travaux récents dans les théories LFG et HPSG initialement issues de la grammaire syntagmatique. Par exemple, la séparation entre la structure de dépendance et la structure de constituant est clairement affirmée en LFG avec la considération des f-structures et c-structures (f pour fonctionnel, c pour constituant), mais pour autant, la c-structure reste, dans les grammaires LFG traditionnelles, dans le cadre de la syntaxe X-barre (ce que nous dénonçons dans [C19]). De même, en HPSG, malgré la considération d'une structure topologique depuis Kathol 1995, une structure syntagmatique de type X-barre est toujours considéré (bien que, comme je le défends dans [A7], elle n'ait plus de motivation théorique).

Nous allons terminer cette section sur l'ordre des mots par des questions computationnelles. Les constructions non projectives, comme on trouve dans la syntaxe de l'allemand ou en français dans le cas des extractions et de l'inversion du sujet, posent des problèmes de traitements. La raison en est la suivante : dans une construction projective, chaque mot est, par définition, dans la projection de son gouverneur, c'est-à-dire qu'il peut être placé par rapport à son gouverneur par un syntagme associant à la dépendance entre lui et son gouverneur une spécification d'ordre, comme proposé dans la Section 3.3. Une grammaire de ce type est équivalente à une grammaire de réécriture hors-contexte et permet donc, comme ces grammaires, une analyse en $O(n^3)$, où n est la longueur des phrases.

Dans [C4], nous proposons, avec Alexis Nasr et Owen Rambow, un analyseur polynomial pour des grammaires permettant de traiter des structures non projectives, à condition de borner le nombre d'éléments placés hors d'une même projection. La même stratégie d'analyse est possible pour la grammaire de l'allemand proposée en [C15]. Je vais expliquer le plus simplement possible et en laissant de côté tous les détails techniques l'idée qui sous-tend cet algorithme d'analyse. L'algorithme est de type CKY (Cocke-Kasutami-Young). Dans le cas projectif, l'idée centrale est la suivante : la distribution d'un segment de la phrase qui est la projection d'une tête lexicale ne dépend que de la nature de cette tête[66]. Pour analyser une phrase, on construit par récurrence sur la longueur des segments tous les segments d'une phrase qui sont potentiellement la projection d'une tête lexicale. A chaque étape de la récurrence, il n'est nécessaire que de garder en mémoire la description de la tête des segments analysés, et ainsi la quantité d'information nécessaire sur chaque segment analysé reste bornée bien que la taille des segments augmente. Comme, pour une phrase de longueur n, il y a n étapes dans la récurrence et que, à chaque étape, il y a moins de RC^2n^2 calculs à faire (où C est le nombre maximum de descriptions d'un segment et R le nombre de syntagmes, sachant qu'il faut combiner deux segments par un syntagme et qu'il y a moins de n^2 couples de segments considérés à chaque étape), l'analyse est en $O(n^3)$ (l'analyse réussit si l'on arrive à construire un segment saturé de longueur n). L'algorithme dans le cas projectif est présenté

[66] Une projection n'est pas nécessairement une projection maximale. La distribution d'une projection dépend donc de la valence passive de la tête (= les positions que peut occuper cette tête), mais aussi de sa valence active non remplie (les dépendants potentiels de cette lexie).

en détail dans la Section 5 de [D1]. Si l'on analyse maintenant une phrase dont la structure n'est pas projective, on ne peut plus se contenter de considérer les segments qui sont des projections. Si l'on prend l'exemple du groupe nominal *la pièce dans laquelle veulent jouer les enfants*, l'analyse va bloquer après la deuxième étape (la construction des segments de longueur ≤ 2), puisqu'on ne peut combiner ni *dans laquelle* et *veulent*, ni *jouer* et *les enfants*. La solution consiste à autoriser la reconnaissance de segments insaturés, c'est-à-dire qui sont la projection d'une tête lexicale à un, deux, trois, … trous près (angl. *gaps*). Mais la distribution d'un segment à k trous ne dépend plus de la seule tête du segment : elle dépend aussi de la description des k éléments manquants. Si l'on borne le nombre maximal de trous dans un segment, l'algorithme reste polynomial en $O(n^{3+2k})$ (vu que le nombre maximal de description d'un segment est en $O(n^k)$) ([A4]).

Pour adapter l'algorithme à une grammaire comme celle de [C15], même dans le cas projectif, la description de la tête ne suffit pas : il faut aussi connaître la nature du constituant topologique ouvert par cette tête. Effet, un verbe qui a ouvert un domaine enchâssé n'a pas la même combinatoire qu'un verbe qui est dans la parenthèse droite du domaine de son gouverneur : le premier peut se combiner avec n'importe lequel de ses dépendants, le deuxième ne peut se combiner qu'avec un dépendant verbal. Ensuite, pour traiter les cas non projectifs (c'est-à-dire les cas où un élément n'est pas dans un des constituants topologiques ouverts par son gouverneur), plutôt que d'enregistrer la description des éléments manquants dans le constituant, on enregistrera la description des éléments susceptibles d'avoir un dépendant hors du domaine de la tête. Ceci optimise l'algorithme, puisqu'un même élément peut avoir plusieurs dépendants hors du domaine de la tête et donc contrôler plusieurs "trous" à la fois. On obtient toujours un algorithme en $O(n^{2k+3})$, où k est maintenant le nombre maximal d'éléments gouvernant un élément émancipé. L'émancipation est un phénomène courant en allemand, mais qui ne peut pas concerner plus d'un ou deux éléments sans réduire nettement l'acceptabilité des exemples. Avec $k = 2$, on obtient une complexité en $O(n^7)$, équivalente aux analyseurs pour les TAG. Or, les TAG ne peuvent non plus traiter plus de deux émancipations, mais, en plus, il existe des exemples on ne peut plus standard comportant une seule émancipation qu'elles ne peuvent traiter (comme *Ich habe das Buch gelesen, das du mir empfohlen hast* 'J'ai lu le livre que tu m'as conseillé' pour laquelle il n'est pas possible d'adjoindre la relative extraposée sur son antécédent).

Les algorithmes de type CKY récupèrent en même temps toutes les analyses d'une phrase ambiguë. Mon intérêt se tourne plutôt vers des algorithmes d'analyse cognitivement motivés, c'est-à-dire incrémentaux, qui permettent de récupérer les analyses les unes après les autres (et de s'arrêter après la première). De tels algorithmes nécessitent en cas d'échec un retour en arrière au dernier choix qui a été fait. On peut rendre de tels algorithmes polynomiaux en mémorisant (ou "tabulant") les morceaux d'analyse abandonnés lors d'un retour en arrière, pour éviter de faire plusieurs fois le même travail. Une partie de la Section 5 de [D1] est consacrée à l'analyse incrémentale et différentes pistes ont été évoquées, dans mes articles, pour le traitement des structures non projectives, comme en [C1] ou en [A3]. Je ne reviendrai pas sur [A3] et les arbres à bulles. En [C1], l'analyse incrémentale est envisagée comme le

traitement d'un flux de dépendances potentielles entre les mots déjà analysés et les mots qui restent à analyser. La projectivité assure que les dépendances introduites par les derniers mots considérés devront être traitées en premier. On peut modéliser le traitement du flux par un analyseur à pile dont les symboles de pile sont les liens potentiels. La non-projectivité se traduit par le fait qu'un lien récent (sur le dessus de la pile) puisse être traité après un lien plus ancien. On peut alors contrôler la non-projectivité par un contrôle sur l'ordre des liens dans la pile. Ces idées, à mon avis prometteuses, ne sont qu'effleurées dans [C1]. Il faut également noter que le flux de dépendances ne semble jamais dépasser 7, ce qui permet de borner la taille de la pile et d'envisager des algorithmes linéaires en temps. Ceci n'a rien d'irréaliste dans la mesure où l'on cherche à modéliser une activité humaine qui est effectuée en temps linéaire (dans un processus de communication ordinaire réussi).

6 La modélisation des collocations

Cette section présente deux articles récents qui portent sur la modélisation des collocations par les fonctions lexicales. Nous allons insister tout particulièrement sur le statut original que confère aux fonctions lexicales l'architecture de GUST et la théorie des signes sur laquelle repose GUST.

Igor Mel'čuk est l'un des premier à avoir distingué nettement, parmi les locutions, les phrasèmes complets comme CASSER LA CROUTE, SE TAPER LA TETE CONTRE LES MURS ou BOIRE LE BOUILLON des semi-phrasèmes comme *tomber de fatigue*, *courir un risque* ou *faire cours*. Les premiers sont des signes profonds élémentaires et par conséquent des entrées du lexique sémantique : aucun des mots de la locution n'a son sens habituel et ne peut être considéré comme un signe profond. Dans les deuxièmes, il en va tout autrement. Bien que l'ensemble forme une expression figée, l'un des mots, ici le nom, a son sens usuel. Les semi-phrasèmes sont donc considérés comme des *collocations* de deux termes : une *base* (ici le nom) choisie librement par le locuteur en fonction de son sens et un *collocatif* (ici le verbe) choisi en fonction de la base de façon arbitraire et contingente pour exprimer un sens ou assurer la réalisation d'une construction syntaxique particulière.

Dans le cadre de la TST, les collocatifs sont répertoriés dans l'entrée lexicale de leur base par l'intermédiaire de fonctions lexicales (Mel'čuk *et al.* 1984-99). Formellement, une *fonction lexicale* [dorénavant FL] est une fonction qui associe un collocatif (ou un dérivé sémantique[67]) à une entrée lexicale et dont l'expression décrit grossièrement la construction syntaxique de la collocation et l'apport de sens du collocatif. Une grande partie des unités de sens apportées par les collocatifs appartiennent à un petit ensemble de sens assez généraux : 'intense', 'bon', 'causer', 'se manifester', 'commencer', etc. En combinant ces sens avec les constructions syntaxiques possibles, on obtient un petit jeu de FL appelées les *FL simples* pour lesquelles un

[67] L'une des caractéristiques notables des FL est d'unifier la description des collocations et des dérivations sémantiques : par exemple, à partir de la base *aimer*, on décrira parallèlement la collocation *aimer à la folie* et la dérivation sémantique *adorer*.

nom conventionnel a été adopté : Magn, Bon, Caus, Manif, Incep, $Oper_i$, $Real_i$, etc. Les *FL complexes*, mettant en jeu plusieurs unités de sens, sont décrites par des combinaisons de FL simples. Enfin, les FL mettant en jeu d'autres sens que les sens généraux des FL simples, sont appelées *FL non standard*, et sont décrites par une paraphrase en langue naturelle contrôlée.

L'encodage classique des fonctions lexicales, élaboré autour d'Igor Mel'čuk depuis plus de 30 ans, n'est pas entièrement satisfaisant : la combinatoire des FL n'est pas complètement formalisée, ni le calcul des caractéristiques sémantiques et syntaxiques d'une collocation en fonction des caractéristiques de la base et de l'expression de la FL. De plus, le statut théorique des FL n'est pas non plus totalement élucidé. Nous allons voir quelles réponses GUST apporte à ces questions.

Igor Mel'čuk et d'autres ont mis en avant le fait que d'un certain point de vue les FL se comportent comme des lexies dont le signifié est vague et le "signifiant" variable. Par exemple, la FL Magn signifie grossièrement 'intense' et se réalise par les expressions *follement*, *hautement*, *comme un singe* ou *comme la lune* selon qu'elle se combine avec AMOUREUX, IMPROBABE, MALIN ou CON. Malheureusement, cette conception des fonctions lexicales n'a jamais reçu de fondements théoriques et elle se heurte au fait que des expressions comme *comme la lune* sont par ailleurs des expressions libres de la langue et que les deux expressions — la valeur de FL et l'expression libre — partagent, à l'exception du sens, un grand nombre de propriétés. Dans l'architecture de GUST, les FL se présentent très clairement comme des signes profonds : les expressions telles que *tomber de fatigue*, *courir un risque* ou *faire cours* seront traitées comme la combinaison de deux signes profonds, une lexie ordinaire (ici le nom) et une fonction lexicale : $\underline{Oper_1} \oplus \underline{\text{FATIGUE}}$, $\underline{Oper_2} \oplus \underline{\text{RISQUE}}$, $\underline{Real_1} \oplus \underline{\text{COURS}}$[68]. La différence avec la description d'une lexie ordinaire est que le signifiant reste sous-spécifié dans la description de la FL et ne peut être spécifié que par la combinaison de la FL avec une base. C'est ce qui justifie le terme de *fonction lexicale* : le signifiant d'une FL est fonction de la base avec laquelle il se combine. Grâce à notre architecture à trois étages, les FL, comme les locutions, sont considérées comme des signes profonds dont le signifiant est un fragment de représentation syntaxique, c'est-à-dire une combinaison de lexèmes, grammèmes et relations syntaxiques. Chacun des éléments qui figure dans la valeur d'une FL est, comme tout élément du signifiant d'un signe profond, le signifié de signes intermédiaires et fait à ce titre l'objet d'une description complète (à savoir la description des signes intermédiaires concernés). De plus, lorsque, comme dans bien des cas, les éléments qui composent la valeur d'une FL apparaissent aussi comme signifiants d'autres signes profonds, la description relative à ces éléments est naturellement mise en facteur et n'a pas à être répétée.

[68] Nous soulignons dorénavant les noms de FL comme nous le faisons pour tous les noms de signes profonds. Pour la définition des FL simples, cf. [C17] ou Mel'čuk *et al.* 1995.

La modélisation des FL par des signes profonds à signifiant variable fonctionne tant qu'on ne veut pas prendre en compte les variations de sens des différentes valeurs d'une FL[69]. Si on veut prendre en compte ces différences de sens, il faut considérer que chaque valeur de FL est le signifiant d'une lexie propre et voir la FL comme un schéma sémantico-syntaxique pour un ensemble de lexies. La FL redevient une fonction lexicale à part entière, c'est-à-dire une fonction associant à une lexie (la base) une autre lexie (le collocatif). Une collocation n'est plus décrite par la combinaison d'une FL et d'une base (f ⊕ L), mais par la combinaison d'une valeur de FL et d'une base : f(L) ⊕ L.

Je vais maintenant présenter brièvement les articles [C17] et [A6] au vu de la modélisation des FL adoptée ici.

Dans [C17], nous nous sommes intéressés, avec Alain Polguère, à donner une meilleure assise formelle à l'encodage des FL et tout particulièrement à l'encodage des FL complexes. Nous avons proposé deux nouveaux encodages. Le premier, appelé *l'encodage explicite*, consiste à décrire explicitement le sens de la FL (c'est-à-dire la combinaison de sens généraux qui forment son sens) et la construction syntaxique que forme la collocation. Par exemple, la collocation *étouffer sa colère* est décrite par le couple :

(15) Contenu sémantique : Caus[1,Minus[Manif[#]]]
Patron syntaxique : V[1,#]

Le contenu sémantique est une formule logique équivalente à la représentation sémantique de '1 cause la diminution de la manifestation de #' où # désigne la base de la collocation (ici COLÈRE) et 1 son premier actant (ici celui qui ressent de la colère). Le patron syntaxique indique que le collocatif est un verbe dont le 1er actant syntaxique (le sujet) est 1 et le 2ème actant syntaxique est #. En un sens, l'encodage explicite d'une FL met en avant son statut de schéma de signe profond en la présentant comme une association entre un sens et un patron syntaxique. La différence avec la description d'une lexie ordinaire est que le signifiant reste sous-spécifié dans la description de la FL et ne peut être spécifié que par la combinaison de la FL avec une base.

Le deuxième encodage proposé dans [C17], l'*encodage algébrique*, est directement inspiré de l'encodage usuel. Cet encodage décrit en une seule formule le contenu sémantique et le patron syntaxique de la FL. Il repose directement sur la combinatoire des FL et des signes profonds en général. Considérons deux FL f et g décrivant des collocations pour lesquelles f(g(L)) peut avoir un sens. Par définition, f(g(L)) est un collocatif de g(L), qui est lui-même un collocatif

[69] Dans la pratique, la plupart des valeurs de FL ne font pas l'objet d'une description propre. Par exemple, on ne voit pas ce qu'on gagnerait à considérer une entrée COMME LA LUNE pour l'intensifieur de CON. On n'aura rien à mettre de plus dans cette entrée que le fait qu'il s'agit d'un intensifieur. Par contre si l'on prend des expressions comme *réussir un examen*, qui est décrite comme $Real_2$ ⊕ EXAMEN, il est clair qu'on voudra une description en soi de RÉUSSIR. Reste qu'il est important de décrire cette expression par une FL et que le choix du verbe *réussir* a quelque chose d'arbitraire (par exemple en anglais, $Real_2$ ⊕ EXAM s'exprime par *to pass an exam*, alors que *to pass* n'est pas une traduction de *réussir* en dehors de ce contexte).

de L. Nous appelons *produit* de f et g la FL, notée f.g, telle que f.g(L) ait les mêmes caractéristiques sémantiques et syntaxiques que f(g(L)) ⊕ g(L). Si l'on voit les FL comme des signes profonds, f.g est la FL qui se comporte comme f ⊕ g, à la différence que f ⊕ g est une expression compositionnelle, alors que f.g est indécomposable dans son signifiant. Dans l'encodage algébrique, la FL décrite en (15) devient $Caus_1$.Minus.Manif où $Caus_1$, Minus et Manif sont des FL simples. Le produit de FL peut être défini à partir de l'encodage explicite assez simplement (cf. [C17]). Il s'agit d'une opération associative et non commutative. Plutôt que de donner la définition formelle de cette opération, je préfère insister sur le fait qu'il s'agit d'une opération naturelle qui correspond à la combinaison linguistique de signes profonds. Pour cette raison, l'encodage algébrique permet de calculer facilement des paraphrases linguistiques et il peut être utilisé pour générer un encodage des FL en langue naturelle (ce qui est utile pour l'enseignement de la langue) ou pour générer des expressions paraphrastiques (ce qui est nécessaire pour la traduction automatique).

Dans [A6], j'explore un autre aspect de l'encodage des FL : le lien entre l'encodage des FL et les définitions sémantiques des lexies. Le problème est le suivant : bien que le sens d'une FL f soit assez vague, le sens d'une collocation f ⊕ L l'est moins. La question est donc de savoir comment se calcule le sens d'une collocation f ⊕ L et comment rendre ce calcul opératoire. Le calcul du sens de f ⊕ L va être effectué à partir du sens de f, donné par la formule encodant f, et par le sens de L, donné par la décomposition sémantique de L. La *décomposition sémantique* d'une lexie est, dans le cadre de la TST, une représentation sémantique substituable au sens de L et contenant des sémantèmes plus simples que ceux de L. Dans les dictionnaires actuellement développés (Mel'čuk *et al.* 1984-99), la décomposition sémantique n'est pas un graphe sémantique, mais une expression linguistique de cette représentation dans un langage contrôlé. L'encodage actuel des décompositions sémantiques n'est quasiment pas structuré ; or, pour rendre le calcul du sens d'une collocation opératoire, il est nécessaire de structurer les décompositions sémantiques et de permettre à (la formule décrivant) la FL de pointer explicitement sur une composante spécifique du sens de sa base. A partir de l'étude critique de l'entrée BLESSURE (Mel'čuk *et al.* 1999), je montre dans [A6] que ces liens ne sont actuellement pas suffisamment explicites et traités de façon hétérogène. En particulier, BLESSURE possède de nombreux types d'intensifieurs portant sur des composantes différentes du sens de la lexie et correspondant généralement à une forme verbale (*grave* vs. *s'aggraver*, *profonde* vs. *se rouvrir*, *mortelle* vs. *succomber à*, *cuisante* vs. *souffrir de*, *vilaine* vs. *s'envenimer*, etc.) pour lesquelles je propose un traitement plus homogène en typant clairement les composantes de la définition lexicographique et en précisant l'encodage des fonctions lexicales.

Pour conclure cette section sur les FL, j'aimerais insister sur l'originalité de la modélisation des collocations des FL. La plupart des approches en sémantique lexicale privilégient le sens de l'analyse et s'intéressent à la polysémie des formes de surface, en cherchant, comme le Lexique Génératif (Pustejovsky 1995) à modéliser les variations de sens d'une même forme. L'approche par FL privilégie au contraire le sens de la synthèse : les expressions de la langue sont envisagées et décrites à partir de leur sens. La modélisation par les FL met en avant la

polymorphie de certains sens, en cherchant à modéliser les variations de forme d'un même sens. Ceci nous amène à considérer des schémas pour des ensembles de signes profonds très variés par leur forme, mais similaires dans leur comportement sémantique et syntaxique.

7 Conclusion

Ce mémoire porte avant tout sur la modélisation mathématique de la langue, c'est-à-dire sur la description formalisée du fonctionnement de différents phénomènes linguistiques. La formalisation n'a pas toujours bonne presse en linguistique, car elle consiste souvent en un encodage brutal et réducteur dans des formalismes complexes reposant sur des présupposés théoriques erronés. Une bonne formalisation doit tout au contraire être fidèle aux concepts sous-jacents et les mettre en lumière. Il faut simplement se donner les moyens (mathématiques) de décrire les choses comme on a envie de les décrire. La mathématisation ne doit pas être un appauvrissement de la pensée.

Il y a de nombreuses façons de formaliser un même concept. Par exemple, un arbre peut aussi bien être vu comme un cas particulier de graphe que comme un cas particulier d'ordre partiel. A rebours, derrière d'apparentes différences de présentation peuvent se cacher des similitudes conceptuelles importantes. La langue est un objet tellement riche et structuré que l'on peut difficilement imaginer un système capable de la simuler sans en adopter plus ou moins les mécanismes. C'est pourquoi j'ai toujours pensé que derrière l'apparente diversité des théories et des modèles se cachent de mêmes idées. Ainsi mon travail de comparaison va-t-il toujours dans le sens de l'unification des approches. C'est pour moi une réelle satisfaction d'avoir pu rapprocher des approches apparemment aussi divergentes que les grammaires génératives, les grammaires transductives de la théorie Sens-Texte ou les grammaires basées sur les contraintes. Ou d'avoir pu rapprocher les arbres de dérivation des TAG et les graphes sémantiques de la TST. Ou encore HPSG et les grammaires de dépendance. Et lorsque deux formalisations ne s'avèrent pas exactement équivalentes, elles donnent chacune un éclairage différent sur un même phénomène et cette différence est riche d'enseignements pour chacune des deux approches.

GUST a été élaboré en privilégiant les choix théoriques sur la formalisation proprement dite. Contrairement à certains modèles linguistiques comme TAG ou les grammaires catégorielles, GUST n'est pas né de l'adaptation d'un formalisme à une théorie linguistique. Au contraire, le formalisme a été développé à partir des données linguistiques et des propriétés de la langue. Par exemple, dans la modélisation de l'extraction, le concept de nucléus a été dégagé, puis un formalisme permettant de traiter les nucléus comme des concepts primitifs a été élaboré, à savoir les arbres à bulles et la grammaire à bulles. De même, dans la modélisation de l'ordre des mots en allemand, les concepts de domaines et de champs se sont avérés centraux et le formalisme a été élaboré pour leur donner la place qui leur revenait.

Le formalisme proprement dit de GUST reste très simple : il s'agit de l'unification la plus simple, c'est-à-dire la superposition de fragments de structures. La richesse du modèle vient avant tout de la richesse des structures manipulées et de leur variété. En particulier, GUST

met en avant les différentes géométries des structures des différents niveaux de représentation. Comme le montre nos structures de dérivation, les combinaisons de signes en GUST ne se limitent pas à une structure hiérarchique d'arbre.

La principale qualité de GUST vient de son architecture qui repose sur une donnée essentielle de la langue, bien connue et pourtant mal prise en compte par les modèles actuels : les unités syntaxiques de la langue — les mots-formes — ne correspondent pas un à un aux unités sémantiques — les sémantèmes. La théorie du signe de Saussure, qui repose implicitement sur une correspondance un à un entre mots-formes et sémantèmes, n'est pas applicable comme telle. Il faut impérativement considérer deux niveaux d'articulation intermédiaires entre le niveau sémantique (le niveau des sémantèmes) et le niveau phonologique (le niveau des phonèmes) : le niveau syntaxique (le niveau des mots-formes ou lexèmes) et le niveau morphologique (le niveau des morphèmes). En effet, d'une part, un mot-forme peut exprimer une combinaison de sémantèmes et être décomposé en autant d'unités signifiantes (les morphèmes) ; et d'autre part, un sémantème peut être exprimé par une combinaison de mots-formes, comme c'est le cas pour les locutions, les prépositions régimes, les auxiliaires, etc. Il faut donc pouvoir traiter séparément la correspondance entre sémantèmes et mots-formes, la correspondance entre mots-formes et morphèmes et la correspondance entre morphèmes et phonèmes. En ajoutant la correspondance entre phonèmes et sons réels, ceci nous conduit à postuler une quadruple articulation de la langue, que nous prenons comme base de la modélisation.

GUST doit beaucoup à la Théorie Sens-Texte dans ses choix théoriques. Néanmoins, GUST n'est pas exactement une nouvelle formalisation de la TST. L'architecture a été simplifiée, en éliminant certains niveaux de représentation, et les interactions entre les règles (les signes en GUST) d'un même module comme de deux modules adjacents ont été éclaircies et enrichies, permettant de simplifier d'autant les règles proprement dites. Enfin, la structure de certaines représentations a été enrichie, notamment la structure syntaxique avec les arbres à bulles et la structure morphologique avec la structure topologique.

GUST est loin d'être un modèle linguistique abouti. Cette étude et les travaux qui l'ont précédée ne sont qu'un début. Mes premiers travaux en modélisation des langues datent d'à peine 6 ans et l'architecture qui sous-tend GUST ne s'est réellement dégagée qu'au cours des 2 dernières années. Dans cette présentation, j'ai plusieurs fois confronté GUST à d'autres approches, mais on ne peut non plus réellement comparer un modèle émergeant comme GUST avec des modèles maintenant bien établis comme HPSG et TAG, qui ont plus de 15 ans d'existence et une communauté de plusieurs dizaines de chercheurs derrière eux. Le fragment de GUST que j'ai présenté est encore bien loin d'avoir une couverture suffisante. Des phénomènes incontournables comme la coordination[70], la négation, la comparaison ou les

[70] Dans [C2], je propose un traitement, par des arbres à bulles, de divers phénomènes de coordination comme la coordination récursive (*Pierre invitera soit Marie et Pierre, soit Jean*), la coordination distribuée (*une assiette et un plat blancs*), la coordination elliptique (*Pierre veut aller à Paris et Marie à Londres*) ou la coordination d'un morceau de valence (*Pierre donne un livre à Marie et un cahier à Jean*). Un fragment de grammaire à bulles

déterminants complexes ne sont pas encore traités. La sémantique de GUST reste très pauvre et certains aspects ne pourront être longtemps négligés comme la sémantique des quantificateurs (les phénomènes de portée) ou la sémantique discursive (les relations de discours). Le rôle de la structure communicative dans le choix de la structure topologique de la phrase (l'ordre des mots et les constituants de surface) et de certaines structures syntaxiques (comme la dislocation ou le clivage) a fait l'objet d'études préliminaires, mais n'est pas encore intégré à l'architecture de GUST et devrait faire évoluer le formalisme des représentations. L'intégration d'un véritable module morphologique (permettant de combiner des morphes pour former des mots) est prévu dans l'architecture de GUST, mais rien n'a encore été développé.

Une autre des faiblesses de GUST est qu'il est pour l'instant, à l'image des modèles par rapport auxquels il se situe (notamment TAG, HPSG et LFG), essentiellement un modèle syntaxique. Or un modèle linguistique ne se résume pas à un modèle syntaxique. Si les grammaires formelles mettent depuis quelques années davantage l'accent sur l'interface sémantique-syntaxe que sur la syntaxe proprement dite, il existe des composantes essentielles de la langue qui sont encore négligées par les modèles actuels. Je pense en particulier à la prosodie et la structure communicative (ou structure informationnelle). Aucun modèle linguistique qui a l'ambition de décrire la langue dans son entier ne met en avant de telles informations à ma connaissance (à l'exception de la TST où elles ont leur place dans l'architecture globale du modèle, mais reste néanmoins moins étudiées). Mêmes des théories linguistiques bien développées comme HPSG y font peu référence (aucune référence ni à la prosodie, ni à la structure communicative dans Pollard & Sag 1994 ou Sag & Wasow 1999). Pourtant décrire une construction comme la dislocation gauche (*Pierre, je ne lui fais pas confiance*) ne peut se limiter à dire que sous certaines conditions syntaxiques un groupe nominal peut être antéposé et coréférent avec un pronom personnel (ici *Pierre* et *lui*). Cette construction induit une prosodie bien particulière (notamment un accent topical sur le groupe nominal antéposé) et nécessite un contexte communicatif précis (le groupe nominal antéposé est un thème proéminent, c'est-à-dire l'objet dont on parle et que de plus on souhaite contraster avec tout objet dont on pourrait dire la même chose). Un modèle linguistique doit contenir les informations que l'on vient de mentionner. Par exemple, les informations communicatives et prosodiques sur la dislocation gauche sont essentielles si l'on veut pouvoir synthétiser cette construction dans un système de génération de textes (sinon comment choisir de faire une dislocation gauche et comment la réaliser correctement, prosodie incluse).

Enfin GUST reste pour l'instant un modèle essentiellement théorique. Son caractère opératoire est évident, mais il est maintenant nécessaire de développer une version électronique de GUST, ainsi que des algorithmes d'analyse et de synthèse. J'aimerais rappeler à ce propos que GUST a été élaboré en cherchant à éviter absolument les ambiguïtés artificielles en associant une unique structure à chaque signe, laquelle peut s'utiliser dans tous

---

correspondant a été écrit, mais non publié. Le principal problème reste pour moi la représentation sémantique de ces phénomènes de coordination.

les emplois du signe. J'ai le sentiment qu'une grammaire de ce type permet d'obtenir des algorithmes beaucoup plus efficaces pour l'analyse comme pour la synthèse. Cela reste évidemment à prouver par la théorie comme par la pratique.

Références

(la liste de mes travaux fait l'objet d'une rubrique spéciale à la suite de celle-ci)

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

Abeillé Anne, Candito Marie-Hélène, 2001, "FTAG: A lexicalized tree Adjoining Grammar for French", in A. Abeillé & O. Rambow, *Tree Adjoining Grammar: Formalisms, Linguistic Analysis and Processing*, CSLI Publications, Stanford, 305-29.

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

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

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

Bech Gunnar, 1955, *Studien über das deutsche Verbum infinitum*, 2nd edition 1983, Linguistische Arbeiten 139, Niemeyer, Tübingen.

Berrendonner Alain, 1991, "Pour une macro-syntaxe", in D. Willems (éd.), *Données orales et théories linguistiques*, Duculot, Paris, 25-31.

Blache Philippe, 2001, *Grammaires de propriétés*, Hermès.

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

Blanche-Benveniste Claire, 1997, *Approches de la langue parlé en français*, Ophrys, Paris.

Bohnet Bernd, Wanner Leo, 2001, "On using a parallel graph rewriting formalism in generation", *Workshop on Natural Language Generation, ACL 2001,* Toulouse.

Bouma Gosse, Malouf Robert, Sag Ivan, 2001, "Satisfying Constraints on Extraction and Adjunction", *Natural Language and Linguistic Theory*.

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

Bresnan Joan, 1982, "The passive in lexical theory", in J. Bresnan (éd.), *The Mental Representation of Grammatical Relations*, MIT Press.

Bresnan Joan, 1998, *Lexical-Functional Syntax*, Blackwell, Oxford.

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

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.

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

Chomsky Noam, 1995, *Minimalist Program*, MIT Press, Cambridge.

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

Deulofeu José, 1989, "Les couplages de constructions verbales en français : effet de cohésion discursive ou syntaxe de l'énoncé", in *Recherches sur le français parlé, 9*, 111-41.

Donohue Cathryn, Sag Ivan, 1999, "Domains in Warlpiri", *HPSG 1999*, Edinburgh.

Dubois Jean *et al.*, 1994, *Dictionnaire de linguistique et des sciences du langage*, Larousse.

Drach Erich, 1937, *Grundgedanken der deutschen Satzlhere*, Diesterweg, Frankfurt.

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

Gerdes Kim, en préparation, *Des grammaires d'arbres lexicalisées pour l'allemand*, Thèse de l'université Paris 7.

Ginzburg Jonatahn, Sag Ivan, 2001, *Interrogative Investigations: The Form, Meaning, Use of English Interrogatives*, CSLI Publications/University of Chicago Press.

Goldberg Adele E., 1995, *A Construction Grammar A pproach to Argument Structure*, Univeristy of Chicago Press.

Hudson Richard, 1984, *Word Grammar*, Blackwell, Oxford.

Hudson Richard, 1990, "Discontinuity", in S. Kahane (éd.), *Grammaires de dépendance, T.A.L., 41:1*, Hermès, 15-56.

Iordanskaja Lidija, Kittredge Richard, Polguère Alain, 1988, "Implementing a Meaning-Text Model for Language Generation", *COLING 1988*.

Iordanskaja Lidija, Mel'čuk Igor, à paraître, "Towards the Notion of Surface-Syntactic Relations", in A. Polguère (éd.), *Collection of papers on dependency in the Meaning-Text framework*, Benjamins, 80 p.

Jackendoff Ray S., 1977, *X' Syntax: A Study of Phrase Structure*, Linguistic Inquiry Monograph, MIT Press.

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

Joshi A ; Levy L., Takahashi L., 1975, "Tree Adjunct Grammars", *Journal of the Computer and System Sciences*, 10 :1, 136-163.

Joshi Aravind, Kallmeyer Laura, 1999, "Factoring Predicate Argument and Scope Semantics: Underspecified Semantics with LTAG", in P. Dekker (éd.), *12th Amsterdam Colloquium Proceedings*, 169-174.

Kaplan Ronald, Bresnan Joan, 1982, "Lexical Functional Grammar: A formal system for grammatical representation", in J. Bresnan (éd.), *The Mental Representation of Grammatical Relations*, MIT Press.

Kasper R., Kiefer B., Netter K. & Vijay-Shanker K., 1995, “Compilation of HPSG to TAG”, *ACL 1995*, 1995, Boston.

Kathol Andreas, 1995, *Linerarization-based German Syntax*, PhD thesis, Ohio State University.

Kay Martin, 1979, “Functional grammars”, *Proc. 5e Annual Meeting of the Berkeley Linguistic Society*, Berkeley, 142-58.

Kay Martin, 1984, “Functional unification grammar: A formalism for machine translation”, *COLING 1984*, Stanford, 75-78.

Kayne Richard, 1974-75, “French relative *que*”, *Recherches Linguistiques II*, 40-61, et *III*, 27-92.

Kroch Anthony, Joshi Aravind, 1986, “Analysing Extrapositions in TAG”, in G. Huck & A. Ojeda (eds.), *Discontinuous Constituents, Syntax and Semantics 20,* Academic Press, 107-49.

Lambek J., 1958, “The mathematics of sentence structure”, *American Mathematical Monthly, 65*, 154-170.

Lecomte Alain, 1992, “Proofnets and dependencies”, *COLING 1992*, Nantes.

Le Goffic Pierre, 1993, *Grammaire de la phrase française*, Hachette.

Martinet André, 1960, *Eléments de linguistique générale* ; 1980, troisième édition revue et corrigée, Armand Colin.

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

Mel’čuk Igor, 1993, *Cours de morphologie générale*, Volume 1 : Le mot, Presses de l’Université de Montréal/CNRS.

Mel’čuk Igor, 1995a, *The Russian Language in the Meaning-Text Perspective*, Wiener Slawistischer Almanach/Škola "Jazyki russkoj kul´tury", Vienne/Moscou, 682 p.

Mel’čuk Igor, 1995b, “Phrasemes in Language and Phraseology in Linguistics”, in M. Everaert, E.-J. van der Linden, A. Schenk, R. Schreuder (éds), *Idioms. Structural and Psychological Perspectives*, Lawrence Erlbaum Associates, Hillsdale, N.J., 167-232.

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

Mel’čuk Igor, 2001, “Grammatical Voice in French (A Short Description in the Meaning-Text Framework)”, in V. Chrakovskij, M. Grochowski, G. Hentschel (éds), *Studies on the Syntax and Semantics of Slavonic Languages (Papers in Honour of Andrzhej Bogusław-ski on the Occasion of his 70th Birthday),* 2001, Universität Oldenburg, 265-290.

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

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.

Milićević Jasmina, à paraître, “Linear Placement of Serbian Clitics: A Description within a Dependency-Based Approach”, in A. Polguère et L. Wanner (eds.), *Topics in Dependency Syntax*, Benjamins, Amsterdam.

Miller Philip, 1999, *Strong Generative Capacity: The Semantics of Linguistic Formalism*, CSLI Publications, Stanford.

Miller Philip, Sag Ivan, 1993, “French clitic movement without clitics or movement”, *LSA Meeting*, Los Angeles.

Moortgat Michael, 1988, *Categorial Investigations: Logical and Linguistics Aspects of the Lambek Calculus*, Foris, Dordrecht.

Morrill Glyn, 1994, *Type Logical Gramar: Categorial Logic of Signs*, Kluwer, Dordrecht.

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.

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.

Pollard Carl, Sag Ivan, 1987, *Information-Based Syntax and Semantics, Volume 1: Fundamentals*, CSLI Lecture Notes 13, Stanford.

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

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

Rambow Owen, 1994, *Natural Language Syntax and Formal Systems*, PhD thesis, University of Pennsylvania.

Rambow Owen, Joshi Aravind, 1992, “A formal look at dependency grammars and phrase-structure grammars, with special consideration of word-order phenomena”, in L. Wanner (éd.), *Current Issue in Meaning-Text Theory*, Pinter, London.

Reape Mike, 1994, “Domain union and word order variation in German”, in J. Nerbonne, K. Netter & C. Pollard (éds), *German in Head-driven Phrase Structure Grammar*, CSLI Lecture Notes 46, Stanford.

Rétoré Christian, 2001, *Logique linéaire et syntaxe des langues*, Habilitation à diriger les recherches, Nantes, 193p.

Sag Ivan, Wasow Thomas, 1999, *Syntactic Theory: A Formal Introduction*, CSLI Publications, Stanford.

Shieber Stuart, Schabes Yves, 1994, “An alternative conception of tree-adjoining derivation”, *Computational Linguistics, 20:1*, 91-124.

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

Tesnière Lucien, 1959, *Eléments de syntaxe structurale*, Klincksieck, Paris.

Van Valin Robert Jr, LaPolla Randy, 1997, *Syntax: structure, meaning and function*, Cambridge University Press.

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., 1987, *A Study of Tree Adjoining Grammars*, PhD thesis, University of Pennsylvania.

Wasow Thomas, 1977, "Transformations and the lexicon", in P. Culicover, T. Wasow, A. Akmajian (éds), *Formal Syntax*, Academic Press, 327-60.

Zaenen Annie, Kaplan Ronald 1995, "Formal devices for linguistic generalizations: West Germanic word order in LFG", in M. Darlrymple, R. Kaplan, J. Maxwell, A. Zaenen, *Formal Issues in Lexical-Functional Grammar*, CSLI Publications.

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

Publications et communications personnelles

(les publications précédées d'un étoile ne sont pas présentées pour l'habilitation)

Articles de revue ou d'ouvrage collectif *(total : 7)*

[A1] Kahane Sylvain, 1998, "Le calcul des voix grammaticales", *Bull. Soc. Ling. Paris, 93:1*, Paris, 325-48.

[A2] 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.

[A3] Kahane Sylvain, 2000, "Extractions dans une grammaire de dépendance lexicalisée à bulles", *T.A.L., 41:1*, 211-243.

*[A4] Kahane Sylvain, à paraître, "The Meaning-Text Theory", *Dependency and Valency, Handbooks of Linguistics and Communication Science*, Berlin/NY: De Gruyter, 32 p.

[A5] Kahane Sylvain, à paraître, "A propos de la position syntaxique des mots *qu-*", in P. LeGoffic (ed.), *Mophèmes en qu-*, Verbum, 44 p.

[A6] Kahane Sylvain, à paraître, "Une blessure profonde dans le DEC : sur le lien entre la définition lexicographique et les fonctions lexicales", à paraître in F. Grossmann & A. Tutin, *Modélisation des collocations*, 11 p.

[A7] Kahane Sylvain, soumis, "HPSG as a True Dependency Grammar: On a More Lexical Treatment of Extraction", 27 p.

Conférences internationales avec actes *(total :19)*

[C1] Kahane Sylvain, 1996, “If HPSG were a dependency grammar ...”, accepté en poster à *HPSG'96* et publié in *Actes TALN’96*, Marseille, 45-49.

[C2] Kahane Sylvain, 1997, “Bubble trees and syntactic representations”, in Becker & Krieger (eds)”, *Proc. 5th Meeting of the Mathematics of Language (MOL5)*, Saarbrücken : DFKI, 70-76.

*[C3] Kahane Sylvain, Mel'čuk Igor, 1997, “Synthèse des phrases à extraction - Aspects sémantiques et syntaxiques”, *Proc. CSSP'97*, Paris, 4 p.

[C4] Kahane Sylvain, Nasr Alexis, Rambow Owen, 1998, “Pseudo-projectivity: a polynomially parsable non-projective dependency grammar”, *Proc. COLING-ACL'98*, Montréal, 646-52.

*[C5] Candito Marie-Hélène, Kahane Sylvain, 1998, “Une grammaire TAG vue comme une grammaire Sens-Texte précompilée”, *Actes TALN'98*, Paris, 40-49.

[C6] Candito Marie-Hélène, Kahane Sylvain, 1998, “Can the derivation tree represent a semantic graph? An answer in the light of Meaning-Text Theory”, *Proc. TAG+4*, Philadelphie, 21-24.

[C7] Candito Marie-Hélène, Kahane Sylvain, 1998, “Defining DTG derivations to get semantic graphs”, *Proc. TAG+4*, Philadelphie, 25-28.

*[C8] Kahane Sylvain, 1999, “Une grammaire de dépendance à bulles pour traiter les extractions”, *Proc.VEXTAL*, Venise, 10 p.

[C9] Kahane Sylvain, Candito Marie-Hélène & de Kercadio Yannick, 2000, “An alternative descriptions of extractions in TAG”, *Proc. TAG+5*, Paris, 115-122.

[C10] Kahane Sylvain, 2000, “How to solve some failures of LTAG”, *Proc. TAG+5*, Paris, 123-128.

*[C11] Kahane Sylvain, 2000, “Des grammaires formelles pour définir une correspondance”, *Actes TALN’2000*, Lausanne, 197-206.

[C12] Kahane Sylvain, 2001, “What is a natural language and how to describe it? Meaning-Text approaches in contrast with generative approaches”, Invited talk, *Computational Linguistics, Proc. CICLing 2001*, Mexico, Springer Verlag, 1-17.

*[C13] Kahane Sylvain, 2001, “A fully lexicalized grammar for French based on Meaning-Text theory”, *Computational Linguistics, Proc. CICLing 2001*, Mexico, Springer Verlag, 18-31.

*[C14] Kahane Sylvain & Polguère Alain, 2001, “Un langage formel d’encodage des fonctions lexicales et son application à la modélisation des collocations”, in B. Daille & G. Williams, *Collocations*, Actes journée Atala, 5 p.

*[C15] Gerdes Kim & Kahane Sylvain, 2001, “A Description of German Syntax Based on a Topological Hierarchy”, *7th Germanic Linguistics Annual Conf. (GLAC’7)*, 1 p.

[C16] Gerdes Kim & Kahane Sylvain, 2001, “Word Order in German: A Formal Dependency Grammar Using a Topological Hierarchy”, *ACL 2001*, Toulouse, 8 p.

[C17] Kahane Sylvain & Polguère Alain, 2001, “Formal Foundations of Lexical Functions”, *Workshop on Collocation, ACL 2001*, Toulouse, 8 p.

[C18] Gerdes Kim & Kahane Sylvain, 2001, “Pas de syntaxe sans prosodie : illustration par l'allemand”, *Journées Prosodie*, Grenoble, 4 p.

[C19] Clément Lionel, Gerdes Kim, Kahane Sylvain, 2002, “An LFG-type Grammar for German Based on the Topological Model”, *LFG 2002*, 7 p.

***Divers** (total : 2)*

[D1] Kahane Sylvain, 2001, “Grammaires de dépendance formelles et théorie Sens-Texte”, Tutoriel, *Actes TALN 2001, vol. 2*, 17-76.

[D2] Kahane Sylvain, 2002, “Du sens au texte : Une modélisation”, *Tangente*, numéro spécial sur les graphes, 3 p. (la revue Tangente est une revue consacrée aux maths, vendue en kiosque et adressée aux lycéens et profs de maths)

***Responsable d’édition** (total : 2)*

*[E1] Kahane Sylvain & Polguère Alain (eds), 1998, *Proc. Workshop on Processing of Dependency-based Grammars, COLING-ACL’98*, Montréal, 150 p.

*[E2] Kahane Sylvain (ed.), 2000, *Grammaires de dépendance, T.A.L., 41.1*, Hermès, Paris.

***Publications mathématiques** (avant 1993) (total : 7)*

*[M1] Kahane Sylvain, 1990, *↑-idéaux de compacts et applications à l’Analyse Harmonique*, Thèse de l’Univ. Paris 6, 81 p.

*[M2] Kahane Sylvain, 1990, “Ensemble de convergence absolue, ensemble de Dirichlet faibles et ↑-idéaux”, *C. R. Acad. Sc. Paris, 310*, 355-357.

*[M3] Kahane Sylvain, 1992, “Opérations de Hausdorff itérées et réunions croissantes de compacts”, *Fundamenta Mathematicae, 141*, 169-194.

*[M4] Kahane Sylvain, 1993, “Antistable classes of thin sets in Harmonic Analysis”, *Illinois J. Math., 37:2*, 186-223.

*[M5] Kahane Sylvain, 1993, “On complexity of sums of Dirichlet measures”, *Annales Institut Fourier, 43:1*, 111-123.

*[M6] Becker Howard, Kahane Sylvain, Louveau Alain, 1993, “Some complete Σ^1_2 sets in Harmonic Analysis”, *Trans. Amer. Math. Soc., 339:1*, 323-336.

*[M7] Kahane Sylvain, 1993, “Finite union of H-sets and countable compact sets”, *Colloquium Math., 65:1*, 83-86.