T.A.L., 1998, vol. 39, n° 2, pp. 57-76

POUR UN MODÈLE STRATIFIÉ DE LA LEXICALISATION EN GÉNÉRATION DE TEXTE

Alain POLGUÈRE*

Résumé - Abstract

Cet article propose un modèle général stratifié de la lexicalisation (calcul et le choix des unités lexicales) en génération de texte. Il est postulé qu'un tel modèle doit être parfaitement intégré à un système global de modélisation des connaissances linguistiques. Ce modèle doit donc être structuré selon les principes d'une théorie linguistique homogène. L'approche linguistique adoptée ici est la théorie Sens-Texte. La section 1 présente le problème de la lexicalisation. La section 2 introduit plusieurs concepts de base permettant de modéliser la lexicalisation dans un cadre Sens-Texte. La section 3 décrit le modèle de la lexicalisation proprement dit.

This paper proposes a general stratificational model of lexicalization (i.e., computing and choice of lexical units) for text generation. It is based on the assumption that such model should be fully integrated into amore general linguistic modeling of natural language. In other words, lexicalization should be structured according to a given linguistic theory. The theoretical framework adopted here is the Meaning-Text linguistic theory. Section 1 presents the problem of lexicalization. Section 2 introduces several key concepts that are needed for a Meaning-Text modeling of lexicalization. Section 3 introduces the lexicalization model proper.

Mots Clefs - Keywords

Génération de texte, lexicalisation, théorie Sens-Texte

Text generation, lexicalization, Meaning-Text theory

---

*Observatoire de linguistique Sens-Texte (OLST), Département de linguistique et de traduction, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal (Québec) H3C 3J7, Canada

(adresse électronique : polguera@ere.umontreal.ca,http://www.fas.umontreal.ca/ling/olst).

La recherche présentée ici est supportée par une subvention du Conseil de Recherches en Sciences Naturelles et en Génie du Canada (OGPO200713) et par une subvention FCAR (97-ER-2741) du Gouvernement du Québec.

Le présent article est une version fortement remaniée d'une communication présentée à la conférence TALN'98. Je remercie vivement L. Danlos, S. Kahane, M. Zock et les membres du comité de lecture de *T.A.L.* pour leurs commentaires, qui m'ont permis (du moins, je l'espère) d'améliorer de façon significative le texte original.

INTRODUCTION

Cet article propose un modèle général de la lexicalisation en génération de texte. Traditionnellement, on entend par *lexicalisation* en génération de texte le processus de calcul et de sélection des unités lexicales du texte (cf., par exemple, Stede 1995). Ce processus a pour pendant celui de grammaticalisation, qui, lui, correspond au calcul des structures syntaxiques et morphologiques du texte à générer. Le modèle de la lexicalisation proposé ici est basé sur une classification systématique des différents types de choix devant être opérés dans le processus de lexicalisation et structurant ces choix de façon logique. La principale caractéristique de ce modèle est qu'il ne repose pas sur l'existence d'une étape particulière de lexicalisation dans le processus de génération ; ce modèle exclut donc le recours à un quelconque« module de lexicalisation ». Il implique au contraire une lexicalisation extrêmement stratifiée, dans laquelle les choix lexicaux se font par étapes successives, en fonction du rôle particulier que chaque unité lexicale doit jouer dans la verbalisation du message que le locuteur (ou la machine) vise à communiquer. L'hypothèse sous-jacente à l'approche que je propose est que la modélisation de la lexicalisation doit être parfaitement intégrée à un système global de modélisation des connaissances linguistiques — grammaticales et lexicales. En conséquence, un tel modèle doit être structuré selon les principes d'une théorie linguistique homogène. L'approche théorique adoptée ici est la théorie linguistique Sens-Texte.

Dans une première section, je vais brièvement présenter le problème de la lexicalisation et son importance en génération de texte. La section 2 introduit plusieurs concepts propres à la lexicalisation, qui me semblent nécessaires pour modéliser celle-ci dans un cadre Sens-Texte. La section 3 décrit l'organisation procédurale de mon modèle. Finalement, la conclusion examine quelques points importants laissés en suspens et présente l'utilisation que je compte faire de ce modèle dans le cadre d'un projet de générateur Sens-Texte en cours de réalisation.

1. LE PROBLÈME DE LA LEXICALISATION EN GÉNÉRATION DE TEXTE

La lexicalisation, c'est-à-dire le processus sélectionnant les unités lexicales d'un texte, est un des champs d'investigation les plus étudiés en ce moment en génération de texte, comme le montrent les nombreuses publications récentes sur ce sujet — par exemple, (Elhadad *et al.* 1997, Nogier & Zock 1992, Stede 1996, Stede 1998, Wanner 1997a et Zock 1996). Une des raisons pour lesquelles la lexicalisation intéresse tant les chercheurs est qu'elle semble se plier plus difficilement que les autres aspects de la génération à la bipartition classique génération profonde — détermination du « quoi dire » — *vs.* génération de surface — détermination du « comment le dire ».

Elhadad *et al.* (1997) identifient trois approches possibles du problème de la lexicalisation, approches qui peuvent être schématiquement représentées de la façon suivante :

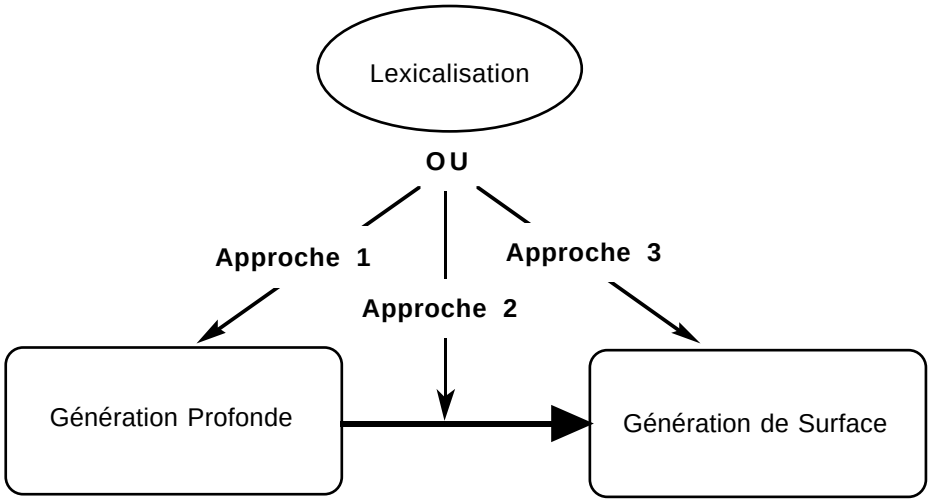

Figure 1. Trois approches de la lexicalisation selon Elhadad *et al.* (1997)

La première approche donne la primauté aux choix lexicaux, opérés pour l'essentiel au stade de la génération profonde ; ils vont contrôler les choix grammaticaux, effectués pour l'essentiel lors de la phase de génération de surface. Cette approche, parce qu'elle est algorithmiquement économique, se trouve implémentée dans de nombreux systèmes de génération. Néanmoins, elle présente de sérieuses limitations. Une des plus graves est sans doute la relative absence de pouvoir paraphrastique résultant d'un tel modèle, basé sur la stratégie « à contenu conceptuel donné, contenu lexical donné ». La rigidité du choix lexical n'est pas nécessairement un problème dans le cadre d'applications où les textes visés appartiennent à un langage technique très contraint, de type sous-langage — (Kittredge & Polguère 1991, Kittredge et *al.* 1994). Cependant, une telle limitation condamne sans appel cette approche si ce que l'on vise, à travers la génération de texte, est la modélisation des processus linguistiques « naturels ».

La seconde approche, intermédiaire, consiste à effectuer les choix lexicaux entre la phase de génération profonde et celle de génération de surface. On trouve une présentation détaillée de l'implémentation de cette approche dans (Elhadad 1992), où est proposée la technique des *contraintes flottantes* (angl. *floating constraints*) pour gérer la propagation des contraintes imposées, notamment, sur les choix grammaticaux par des choix lexicaux effectués antérieurement. Selon Elhadad *et al.* (1997), cette approche est la meilleure. C'est effectivement le cas, si le module de réalisation ne possède pas de véritable pouvoir paraphrastique. (Je reviendrai sur ce point dans un instant.)

La troisième approche consiste à implémenter les choix lexicaux lors de la génération de surface, une fois construit l'arbre syntaxique de la phrase. Cela revient à laisser les choix grammaticaux imposer leurs contraintes sur les choix lexicaux. C'est une approche qui donne aux contraintes de forme (associées à la grammaire) la primauté sur les contraintes de contenu (associées au lexique). Or, on sait bien que les unités lexicales, les lexies, de la langue ne sont pas des entités passives, dont on fait ce qu'on veut et qui se soumettent aux contraintes de la structure syntaxique dans laquelle on les insère. Bien au contraire, les

lexies contrôlent, régissent en grande partie la structure de la phrase. En conséquence, cette troisième approche n'est pas véritablement viable en génération si elle est implémentée de façon aussi radicale.

Même si la seconde approche présentée ci-dessus (et la technique de propagation des contraintes flottantes, qui permet son implémentation) est séduisante *a priori*, je pense qu'elle fait faire aux modules pré-linguistiques du générateur des tâches qui ne leur appartiennent pas. Je propose une quatrième approche, qui est en fait une variante de la troisième, en ce sens qu'elle situe toutes les véritables opérations de lexicalisation au niveau du module de génération de surface. Cette quatrième approche, en effet, consiste à repousser au maximum les choix lexicaux véritables tout en leur reconnaissant un rôle prédominant dans la dynamique de construction de la phrase. Elle présuppose une modélisation extrêmement stratificationnelle de la langue et du processus de génération de surface, c'est-à-dire une modélisation basée sur de multiples niveaux intermédiaires de représentation des énoncés[1]. Même si elle pourra sembler algorithmiquement plus coûteuse, cette approche me paraît être celle qui colle de plus près à la réalité de la langue dans la mesure où il n'existe pas de module de lexicalisation clairement identifiable dans la langue, comme je pense pouvoir le montrer dans ce qui suit. Il est donc justifié d'explorer l'approche que je propose dans une perspective de recherche dont la finalité est d'utiliser la génération comme outil d'étude des connaissances et des processus linguistiques.

On pourrait objecter qu'un programme informatique n'a pas besoin de reproduire le comportement humain pour le simuler. Un système d'intelligence artificielle peut même être beaucoup plus performant si justement il utilise les caractéristiques de traitement de l'information propres à la machine, plutôt que de tenter de coller au traitement fait par l'humain, dans toute sa complexité et son indétermination. On connaît le cas de l'ordinateur Deep Blue de IBM, qui s'est singularisé en battant de façon tout à fait irrespectueuse le champion du monde d'échecs Garry Kasparov. La technique de jeu utilisée par Deep Blue est assez « inhumaine » puisqu'elle est basée sur un calcul systématique, ultra-rapide, d'un maximum de développements de jeu possibles à partir d'une configuration d'échiquier donnée. Cette stratégie permet de construire des machines qui vont certainement, dans un futur très proche, battre de façon systématique les plus grands champions d'échecs. Cependant, Deep Blue est de peu d'intérêt si on veut l'utiliser comme outil de modélisation des connaissances et stratégies mises en œuvre par un joueur d'échecs

[1] Il existe bien entendu d'autres classifications des différentes approches de la lexicalisation, mais, en général, elles recoupent en partie celle de Elhadad *et al.* (1997). Wanner (1997a) propose une classification plus fine, qui introduit notamment ce que l'auteur appelle *lexical choice interwoven with syntactic choice* (choix lexical entremêlé au choix syntaxique). Ce type de lexicalisation correspond en fait très directement, d'un point de vue procédural, à celui que je présente ici. Voir De Smedt (1990), pour un exemple représentatif.

humain[2]. De la même façon, un générateur de texte qui fait une lexicalisation suivant la première approche décrite ci-dessus pourra, dans certains contextes d'application, être très efficace. Cependant, le module de réalisation linguistique (le générateur de surface) qu'incorpore un tel système ne peut en aucun cas être considéré comme un modèle de la langue. La réutilisation des données qu'il possède sur la langue dans d'autres contextes de génération de texte sera problématique et leur utilisation pour d'autres tâches de traitement automatique de la langue (traduction, enseignement assisté, etc.) a de fortes chances d'être une perte de temps pure et simple.

Ce qui est souvent problématique en traitement automatique de la langue, c'est que l'on ne va pas nécessairement s'apercevoir de la raison profonde de certaines difficultés rencontrées. On va parler de problèmes particuliers propres à la langue, sans se rendre compte que ces problèmes sont en fait parfois générés par les concepteurs de systèmes eux-mêmes. Donc, dans la pratique, il faut considérer avec beaucoup de circonspection la solution apportée au problème de la lexicalisation dans un système de génération qui ne ferait même pas la différence entre, par exemple, les collocations (*chute* ***brutale***, ***faire*** *une chute*, …) et les expressions libres (*un garçon brutal, faire un gâteau*, …). L'ingénierie d'un tel système peut-être d'un grand intérêt, mais sûrement pas le modèle de la langue qu'il propose.

Pour conclure ces remarques, il convient de noter que, bien que les raisons profondes qui justifient ma démarche sont de nature théorique, la modélisation de la lexicalisation qui est proposée ici est tout à fait utilisable dans des applications particulières de la génération de texte. C'est notamment ce qui a été tenté, de façon très expérimentale il est vrai, dans des systèmes comme GOSSIP et LFS (Iordanskaja & Polguère 1988, Iordanskaja *et al.* 1991, Iordanskaja *et al.* 1996)[3].

Je vais prendre comme point de départ de mon exposé un postulat assez fort, qui ne fera sans doute pas l'unanimité :

> un modèle général de la lexicalisation en génération de texte doit être enchâssé dans une approche théorique linguistique donnée.

En effet, il me paraît presque impossible de « théoriser » la lexicalisation sans le faire à partir d'une approche linguistique particulière, dans la mesure où l'implémentation de la lexicalisation est un problème procédural et où la structure d'un système procédural est nécessairement héritée de la structure des connaissances — ici, du modèle linguistique — qu'il manipule. Le modèle présenté dans cet article trouve ses racines dans

---

[2] On trouvera une très bonne analyse de la façon de jouer de Deep Blue et de son « intelligence du jeu » dans (Campbell 1997), texte écrit par un des concepteurs de cette machine.

[3] Ces deux systèmes ont été conçus à partir d'une implémentation de modèles Sens-Texte — dictionnaire et grammaire. GOSSIP visait la génération de rapports de sécurité en anglais sur l'activité de systèmes d'exploitation. LFS visait quant à lui la génération bilingue (anglais et français) de rapports statistiques sur l'économie canadienne.

la théorie Sens-Texte — (Mel'čuk & Polguère 1987, Mel'čuk 1997, Polguère 1998 et Wanner 1996, 1997b) ; il hérite donc directement des choix théoriques faits en linguistique Sens-Texte. Néanmoins, on peut s'attendre à ce que nombre de concepts introduits ici trouvent leur transposition naturelle dans d'autres approches, pour lesquelles doit être développé un modèle spécifique de la lexicalisation.

2. BASES CONCEPTUELLES D'UNE LEXICALISATION SENS-TEXTE POUR LA GÉNÉRATION

2.1. Quelques définitions préliminaires

Le concept de lexicalisation, souvent aussi désigné par le terme de *choix lexical*, est assez problématique et il convient d'analyser ses présupposés. Si l'on parle de lexicalisation en génération de texte, c'est que l'on présuppose la chose suivante : à au moins une étape du processus de génération, le générateur doit faire des choix ayant pour conséquence l'utilisation dans le texte à venir d'une ou plusieurs lexies que les autres choix possibles ne devraient, normalement, pas autoriser. La lexicalisation serait donc ce sous-processus du processus de génération spécialisé dans la détermination des lexies apparaissant dans le texte. Il est tout à fait possible de considérer, comme on le fait souvent, que le processus de lexicalisation est en fait multiple et s'effectue à plusieurs étapes du traitement, et c'est ce que je ferai ici. La plus importante caractéristique du modèle que je propose est sans doute qu'il ne repose pas sur l'existence d'un module unique de lexicalisation et sur la centralité du concept de choix lexical. En effet, il serait plus approprié de parler ici de calcul lexical, dans la mesure où l'introduction d'une lexie sera le résultat d'un calcul impliquant plusieurs niveaux de choix, qui ne sont pas nécessairement des choix entre lexies. Je reviendrai bien entendu plus loin sur ce point crucial.

Dans la suite de ma présentation, je vais souvent employer le terme de *message*. Je l'utilise comme un terme technique et il est donc important de préciser son sens pour éviter toute confusion avec ses nombreuses autres acceptions (en langue générale, en théorie de l'information, en sémiotique, etc.).

> Dans le contexte de la génération de texte, je considérerai qu'un message est un contenu informationnel devant être exprimé linguistiquement par une phrase.

Un message sera donc (i) de l'information, (ii)structurée pour la communication et (iii) destinée à être communiquée grâce aux ressources lexicales et grammaticales de la langue dans un segment du type phrase. Soit les trois phrases suivantes :

(1) a. *Le cours du dollar a chuté brutalement vers 13h00.*
b. *Le dollar a fait une chute brutale vers une heure de l'après-midi.*
c. *Aux alentours d'une heure de l'après-midi, on a observé une dégringolade du dollar.*

Ces phrases sont des paraphrases linguistiques. C'est-à-dire que tout locuteur du français, en se basant exclusivement sur ses connaissances de

la langue, peut dériver (1b,c) de (1a), (1a,c) de (1b) et (1a,b) de (1c). Bien entendu, il est hors de doute que de fines nuances sémantiques et stylistiques existent entre ces phrases (par exemple, (1c) est stylistiquement, du fait de l'utilisation du terme *dégringolade*, légèrement moins neutre que les deux autres phrases) ; cependant, ces nuances sont suffisamment négligeables pour être ignorées ici.

On dira que (1a-c) expriment le même message. Ce message, quel est-il ? Il est organisé autour d'une bipartition communicative, souvent appelée *opposition thème-rhème*. Le premier tronçon d'information est ce que le message présente comme étant l'information véritablement communiquée — le rhème :'baisse importante et rapide vers 13h00'. Le second tronçon d'information est ce à propos de quoi une information est communiquée — le thème :'cours du dollar'. Le rhème est ici lui-même structuré de façon plus fine en un fait principal — la baisse proprement dite —, des caractéristiques de ce fait — la baisse est importante et rapide — et un paramètre de localisation temporelle — heure de l'événement. Il est bien entendu possible de structurer de façon toute différente le même contenu informationnel : par exemple, on pourrait signaler que la baisse importante et rapide qui a eu lieu vers 13h00 était celle du dollar. Dans ce cas, on a affaire à un message différent, où la structure thème-rhème est inverse de celle de notre premier message.

En fonction de ce qui vient d'être dit, on peut *grosso modo* décrire de façon un peu plus formelle le message exprimé par (1a-c) dans le tableau suivant, où le gras met en évidence les composantes de sens centrales dans le contenu informationnel en question :

INFORMATION COMMUNIQUÉE [= RHÈME]
- Fait principal :
  - 'baisse **baisse** du cours du dollar'
- Caractéristiques :
  - 'baisse **importante**'
  - 'baisse **rapide**'
- Paramètre :
  - '**moment** approximatif de la baisse = 13h00'

… À-PROPOS DE [= THÈME]
- '**cours** du dollar'

On voit que ce qui différencie un message d'un simple contenu informationnel, c'est le fait qu'il est communicativement organisé, c'est-à-dire que sa structure reflète la façon dont l'information doit être présentée dans la phrase. Calculer un message, ce n'est donc pas seulement sélectionner quelle information doit être transmise, c'est aussi trouver comment elle doit être « emballée » par la phrase, pour que celle-ci s'insère dans un contexte pragmatique et textuel (dans le cas de la génération de textes véritables) de façon cohérente. On notera, et c'est un fait essentiel, que le message doit être représenté au moyen d'un nombre restreint d'éléments sémantiques de base, associés à un sous-ensemble du lexique de la langue dans laquelle le texte doit être produit. Le « langage » du message n'est donc pas une *interlingua* — un formalisme unique de représentation dont on peut dériver des textes dans des langues

différentes — puisque son « lexique » est lié à celui d'une langue particulière.

Je vais prendre comme point de départ de la construction de mon modèle les deux axiomes suivants :

> Axiome 1 : La représentation formelle d'un message est ce qui est donné en entrée au module linguistique (*grosso modo*, la grammaire et le dictionnaire) du système de génération.
> Axiome 2 : Si on utilise une modélisation linguistique de type Sens-Texte, qui possède un moteur de paraphrasage, aucun choix lexical n'est nécessaire au niveau du calcul du message.

Une conséquence de ces deux axiomes de départ sur le modèle proposé est que le calcul d'un message linguistique se ramène en fait à un calcul de configurations de sens, sans considération pour les lexies qui seront finalement utilisées pour exprimer ces sens.

Bien entendu, une telle approche n'est possible qu'avec des modèles linguistiques ayant un réel pouvoir paraphrastique. En gros, ce sont des modèles linguistiques dont la fonction est de décrire et implémenter tous les choix qui s'offrent au locuteur, et à la machine, pour exprimer un sens donné. Le modèle linguistique ne doit pas seulement nous dire si c'est A ou B qui est le bon choix, à une étape donnée du traitement. Il doit nous dire que A_1, A_2, A_3, etc. sont toutes les options valides s'offrant à nous. Si le modèle linguistique a cette capacité, on peut lui « faire confiance » : il trouvera toujours le moyen d'exprimer un message bien formé au moyen d'une phrase grammaticale, car c'est exactement ainsi que nous fonctionnons, en tant que locuteurs. Encore une fois, tout ce que j'avance ici est du niveau de l'axiomatique et je ne prétends pas pouvoir le démontrer autrement que par la pratique (voir les remarques à ce sujet dans la conclusion).

2.2. Types de choix impliqués dans la lexicalisation

Lorsque l'on cherche à établir une distinction entre les différents types de choix lexicaux à modéliser en génération, on se limite souvent à l'opposition entre le choix des lexies appartenant aux classes ouvertes (noms, verbes, adjectifs et adverbes) et celui des lexies appartenant aux classes fermées (pronoms, conjonctions, prépositions, etc.). Je pense, au contraire, que la typologie des choix doit reposer sur des critères avant tout sémantiques. Il s'agit de savoir quel type de sens expriment les lexies en question. En fonction de tels critères, on peut postuler deux types majeurs de lexies : les lexies profondes et les lexies de surface.

> Les lexies profondes de la langue possèdent au moins les trois caractéristiques suivantes :
> **1** Elles sont très nombreuses et forment le corps du lexique.
> **2** Leur sens peut être décrit au moyen d'une définition analytique.
> **3** Du fait de la nature de leur sémantisme, ce sont les unités lexicales qui comptent le plus, du point de vue du locuteur, et donc du point de vue de la génération.

Les lexies profondes contenues dans les exemples (1a-c) sont[4] :

(1a) → CHUTER**3**, COURS**III.1**, DOLLAR, HEURE**2**, TREIZE**2**, VERS**1 4**
(1b) → APRÈS-MIDI, CHUTE**I(B)5**, DOLLAR, HEURE**2**, UN**I.4**, VERS**1 4**
(1c) → APRÈS-MIDI, ⌜AUX ALENTOURS DE⌝, DÉGRINGOLADE, DOLLAR, HEURE**2**, UN**I.4**

Apparaissent bien entendu ici des lexies appartenant aux classes grammaticales ouvertes : des noms (COURS**III.1**, DOLLAR, HEURE**2**, …) et un verbe (CHUTER**3**). Mais sont aussi considérées comme lexies profondes certaines lexies appartenant à des classes grammaticales fermées : ici, des numéraux (TREIZE**2** et UN**I.4**), une préposition (VERS**1 4**) et une locution prépositionnelle (⌜AUX ALENTOURS DE⌝). Il n'y a donc pas de recoupement exact entre la notion de classe grammaticale ouverte et celle de lexie profonde. De plus, cette dernière notion n'est pas non plus en correspondance exacte avec celle de lexie pleine : toutes les lexies profondes sont des lexies pleines, mais l'inverse n'est pas vrai. Ainsi, BRUTALEMENT [*chuter brutalement*] est clairement une lexie pleine. Elle n'apparaît cependant pas dans les listes ci-dessus car elle ne répond pas à la deuxième caractéristique des lexies profondes : son sens est difficilement analysable, autrement que par un sens très général comme ‘vite’. On verra plus bas que nous sommes ici en présence d'un cas particulier de lexie de surface : une lexie collocationnelle.

En résumé, les lexies profondes ont normalement une origine directe dans la représentation du message ; cependant, certaines lexies, comme ici BRUTALEMENT (dont le statut sera examiné plus bas), ne sont pas des lexies profondes bien qu'elles expriment directement une composante sémantique du message.

Alors que les lexies profondes sont en quelque sorte des lexies « de base », on peut se représenter les lexies de surface comme étant des lexies de « seconde classe ».

Les lexies de surface de la langue sont de deux types :
1 les lexies grammaticales, qui appartiennent à un petit sous-ensemble du lexique et sont directement liées à la grammaire de la langue (articles, prépositions régies, auxiliaires, etc.) ;
2 les lexies collocationnelles, qui expriment des significations collocationnelles (intensificateurs typiques, verbes supports, etc.) et sont en quelque sorte des lexies profondes « dégénérées ».

Les lexies grammaticales présentes en (1) sont :

(1a) → $\text{AVOIR}_{\text{Aux}}$, DE (cf. *du*), LE_{Art}
(1b) → $\text{AVOIR}_{\text{Aux}}$, DE, LE_{Art}, UN_{Art}
(1c) → $\text{AVOIR}_{\text{Aux}}$, DE, LE_{Art}, $\text{ON}_{\text{Pron. impers.}}$, UN_{Art}

4 Pour référer à des lexies précises, j'utilise la numérotation du *Petit Robert*. Les coins relevés servent à indiquer une locution (cf. ⌜AUX ALENTOURS DE⌝).

Il est important de remarquer que les sens exprimés par de telles lexies sont d'une nature très particulière. Contrairement aux sens des lexies profondes, ils ne peuvent recevoir de définitions paraphrastiques. Comment paraphraser le sens de l'article défini/indéfini, ou, pire, celui de la préposition régie, qui ne fait qu'introduire un complément ? Puisqu'elles relèvent du système grammatical de la langue, il est très difficile de considérer les lexies grammaticales comme étant l'expression directe d'un but communicatif du locuteur, même lorsqu'elles ont une signification véritable (défini/indéfini, participation à l'expression du temps grammatical, etc.). Cette caractéristique a des implications directes sur le modèle que je vais proposer dans la section 3.

Les lexies collocationnelles présentes en (1) sont :

(1a) → (chuter)BRUTALEMENT
(1b) → (chute) BRUTAL**4**, FAIRE^{1}**II.4** (une chute)
(1c) → OBSERVER**II.4** (un phénomène)

Le concept de lexie collocationnelle va permettre de prendre en compte de façon propre la lexicalisation impliquée dans l'expression des significations collocationnelles.

Les significations collocationnelles ont les trois propriétés suivantes :

1 elles s'expriment normalement par des collocations[5] ;
2 elles forment un ensemble très restreint de significations quasi universelles (‘intensification’, ‘causation’, ‘avoir lieu’, etc.) ;
3 elles sont constamment exprimées dans les textes et c'est leur bonne lexicalisation qui donne aux textes leur caractère idiomatique.

Une des innovations les plus importantes introduites par la théorie Sens-Texte est le concept de fonction lexicale syntagmatique, qui permet de décrire de façon rigoureuse les phénomènes collocationnels dans le dictionnaire de la langue. Les fonctions lexicales syntagmatiques sont en quelque sorte des lexies généralisées exprimant une signification collocationnelle. Elles doivent être considérées comme étant un cas particulier de lexies profondes**.** Je ne peux pas entrer ici dans le détail de la présentation des fonctions lexicales (voir les textes de présentation de la théorie Sens-Texte mentionnés à la section 1). Voici toutefois comment se modélisent, au moyen des fonctions lexicales, les quatre collocations mentionnées ci-dessus.

---

[5] Une expression linguistique *AB* (par exemple, *chute brutale*) est une collocation si : (i) elle est la combinaison de deux lexies A (ici, CHUTE) et B (ici, BRUTAL), (ii) le sens ‘AB’ est égal à ‘A’ + ‘C’ (ici, ‘rapide**I.6**’) et (iii) ‘C’ ≠ ‘B’ (‘rapide **I.6**’≠ ‘brutal’). A, l'élément de la collocation qui retient son sens et qui conditionne le choix de l'autre élément, est appelé *base de la collocation*. Pour une description détaillée du concept de collocation, voir (Mel'čuk 1995).

Fonction lexicale **Magn** = modificateur d'intensification :

Magn‘vitesse’(*chute**I(B)5***) = *brutale**4***
Magn‘vitesse’(*chuter**3***) = *brutalement*
[Note: on est ici en présence d'une intensification portant sur la vitesse de déroulement d'un phénomène, plutôt que sur son amplitude, sa fréquence, etc.]

Fonction lexicale **Oper** = verbe support vide prenant la base de la collocation comme premier complément :

Oper$_1$(*chute**I(B)5***) =*faire* [ART ~]
Oper$_0$(*chute**I(B)5***) = **spéc.** [*on*] *observer* [ART ~]

Le cas de cette dernière collocation est un peu particulier. En fait, toute lexie L du type ‘phénomène’ est compatible avec la construction *On observe L*, pour des textes se rapportant à certains domaines de spécialité (ici, l'économie).

Il convient de noter qu'il existe un autre type de fonctions lexicales : les fonctions lexicales paradigmatiques (par exemple, **Syn** pour la synonymie, **V$_0$** pour la dérivation verbale, **S$_i$** pour le nom du i-ème actant prototypique de la lexie, etc.). Ces fonctions lexicales jouent notamment un rôle très important dans les règles de paraphrasage. Néanmoins, pour simplifier mon exposé, je vais me permettre de les ignorer ici. Dans ce qui suit, tout ce qui est dit à propos des fonctions lexicales concerne en fait spécifiquement les fonctions lexicales syntagmatiques.

3. MODÈLE DE LA LEXICALISATION PROPREMENT DIT

3.1. Niveaux de transition impliqués

Je vais maintenant présenter le modèle de la lexicalisation proprement dit, en indiquant comment les concepts introduits dans la section 2 permettent de structurer le processus de lexicalisation.

Avant de commencer la présentation du modèle de la lexicalisation, il convient de faire une mise au point très importante :

> Le modèle que je propose fonctionne exclusivement à l'échelle de la phrase, ce qui est une conséquence directe du fait que, selon moi, la lexicalisation peut être effectuée entièrement au stade de la génération de surface. Je pense cependant que, dans un véritable générateur « intelligent », le module de génération de surface doit avoir accès aux données extralinguistiques pour effectuer/peaufiner ses choix. On peut trouver la présentation de stratégies de ce type dans (McKeown *et al.*1995, Nicolov *et al.* 1997, Zock 1996). Je reviendrai sur ce point dans la conclusion.

Comme il a été mentionné dans la section 1, ma modélisation se caractérise par une extrême stratification du processus de génération, qui va de pair avec la stratification de la génération de surface basée sur

l'approche Sens-Texte. Plus précisément, ce modèle de la lexicalisation est basé sur quatre niveaux de représentation de la phrase : le niveau conceptuel, le niveau sémantique, le niveau syntaxique profond et le niveau syntaxique de surface. Le recours à quatre niveaux implique que la lexicalisation s'effectue à travers trois phases distinctes de transition. Chaque transition va être désignée par son niveau source : la transition conceptuelle (du niveau conceptuel au niveau sémantique), la transition sémantique (du niveau sémantique au niveau syntaxique profond) et la transition syntaxique profonde (du niveau syntaxique profond au niveau syntaxique de surface).

Je vais maintenant brièvement décrire la nature formelle de chacun des quatre niveaux de représentation impliqués. Ensuite, dans la section 3.2., je présenterai une typologie des différents choix liés à la lexicalisation qui sont opérés à chacune des trois phases de transition.

Le niveau conceptuel de représentation correspond à la modélisation d'un état du monde donné sous forme de graphe conceptuel (au sens très large de ‘graphe formé de noms de concepts connectés entre eux’). C'est ce dont la phrase parle, mais ce n'est pas ce que la phrase dit. En d'autres termes, ce n'est pas une représentation du message linguistique encodé dans la phrase. Le contenu informationnel du niveau conceptuel n'est d'ailleurs pas nécessairement limité au domaine de la phrase. Les caractéristiques formelles de ce niveau de représentation varient avec les différentes applications de la génération et il n'est pas utile d'en donner ici un exemple. Notons de plus qu'une représentation conceptuelle peut être ou non une interlingua. C’est au concepteur du système de génération de le décider. De ce point de vue, un générateur censé modéliser de façon réaliste la génération humaine devrait sans doute utiliser des formalismes différents de représentation conceptuelle selon qu'il modélise l'activité linguistique d'un locuteur complètement monolingue ou celle d'un locuteur bilingue, dans une situation de *code switching* potentiel[6].

Le niveau sémantique correspond à la représentation du message linguistique, la représentation sémantique proprement dite. Il est donc du niveau de la phrase. Un exemple de représentation sémantique est donné dans la Figure 2 ci-dessous, qui est une représentation formelle du message exprimé par les exemples (1a-c). Le formalisme utilisé ici est celui des réseaux sémantiques de la théorie Sens-Texte (Polguère 1997). Ce niveau de représentation dépend de la langue mais ne met pas en jeu les lexies elles-mêmes. Il ne contient que la détermination des sens de base à exprimer, ainsi que l'emballage communicatif (structure thème-rhème, etc.) que la phrase devra refléter[7].

---

[6] Le terme *code switching* (‘changement de code’) réfère au processus de passage instantané d'une langue à l'autre dans le discours de locuteurs bilingues (ou, plus généralement, multilingues). Ce qui est intéressant, c'est que ce passage peut s'effectuer à l'intérieur même d'une phrase, produisant une phrase à la structure hybride.

[7] Le lecteur intéressé à en savoir plus sur les représentations utilisées dans les figures qui vont suivre est invité à se reporter aux textes d'introduction sur la théorie

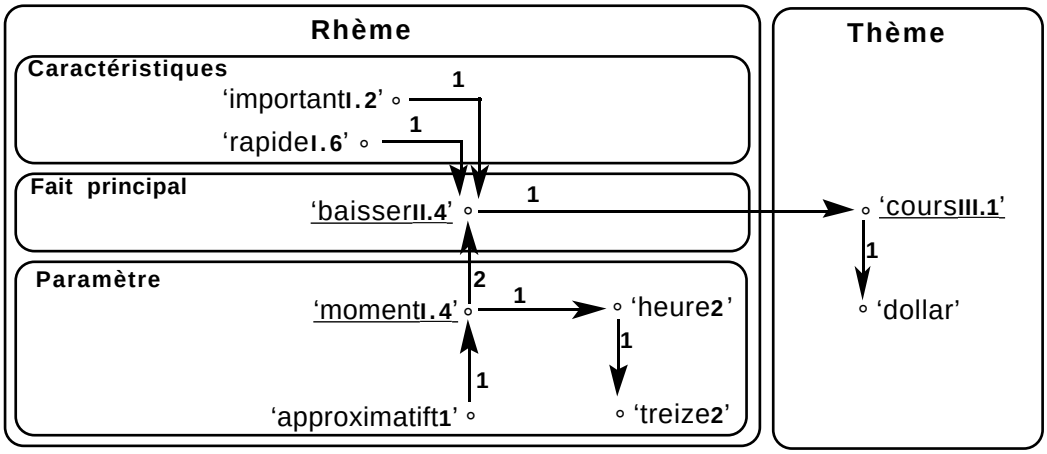

Figure 2. Réseau sémantique correspondant au message exprimé en (1)

Le niveau syntaxique profond correspond à la représentation syntaxique profonde de la phrase à générer. Cette représentation est, formellement, un arbre de dépendances syntaxiques où apparaissent les lexies profondes de la future phrase ainsi que les différentes fonctions lexicales, qui sont, je le rappelle, des cas particuliers de lexies profondes. La Figure 3 ci-dessous est la représentation syntaxique profonde de la phrase (1a) (= *Le cours du dollar a chuté brutalement vers 13h00*.).

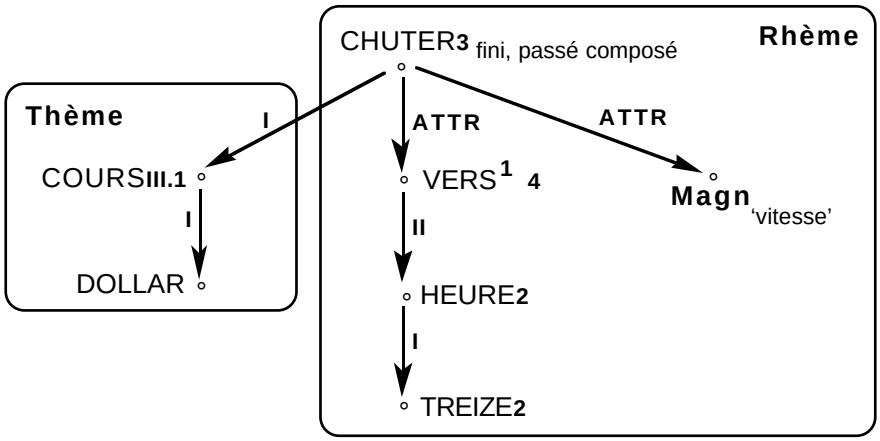

Figure 3. Représentation syntaxique profonde de (1a)

Le niveau syntaxique de surface de la phrase, comme le niveau précédent, est basé sur des représentations qui sont formellement des arbres de dépendances syntaxiques. Ceux-ci contiennent cependant toutes les lexies qui vont être utilisées dans la phrase. C'est-à-dire que sont présentes à ce niveau non seulement les lexies profondes, mais aussi les lexies de surface (grammaticales et collocationnelles). Une fois ce niveau

---

Sens-Texte mentionnés dans la section 1. Il ne sera fait mention ici que des aspects pertinents pour le problème spécifique de la lexicalisation.

atteint, toutes les tâches liées à la lexicalisation ont été effectuées. La Figure 4 ci-dessous est la représentation syntaxique de surface de (1a).

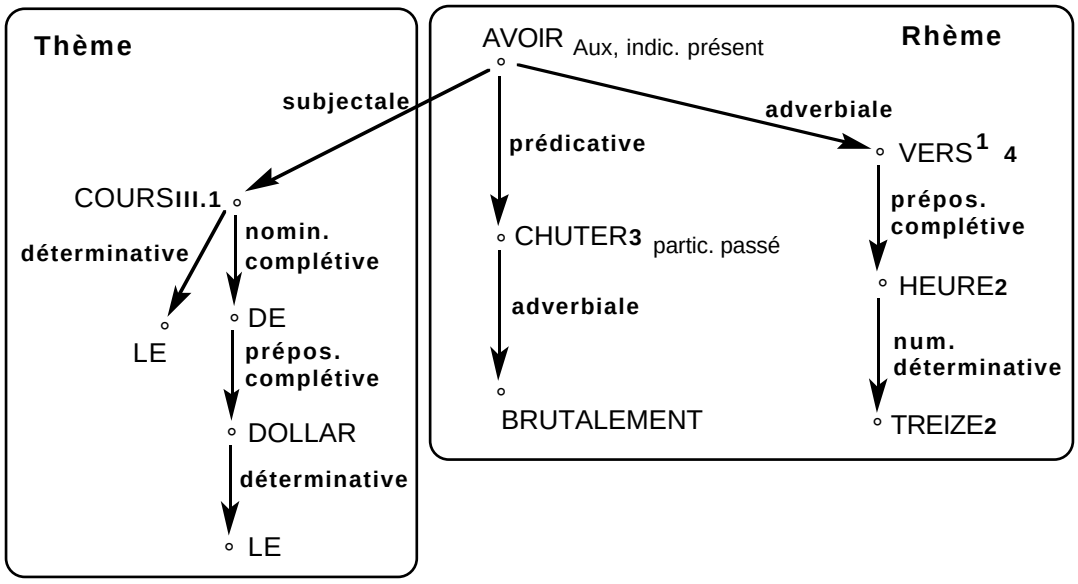

Figure 4. Représentation syntaxique de surface de (1a)

3.2. Typologie des choix opérés à chaque phase de transition

À ce stade de l'exposé, il devient possible de montrer comment s'enchaînent les différents choix linguistiques qui vont permettre de modéliser la lexicalisation en génération de texte. Pour ce faire, je vais présenter une typologie des choix opérés, typologie basée, premièrement, sur la phase de transition concernée et, deuxièmement, sur le type d'unité linguistique sélectionné[8]. Dans ce qui suit, j'utilise les conventions d'écriture suivantes :

Chaque sélection d'une unité linguistique donnée U à un niveau donné (= niveau sémantique, niveau syntaxique profond ou niveau syntaxique de surface) correspond à un des trois cas de figure suivants :

$\rightarrow$U = la sélection de U est essentiellement déterminée par des contraintes provenant du niveau source de la transition ;
$\rightarrow$U$\leftarrow$ = la sélection de U est déterminée par des contraintes provenant du niveau source et du niveau cible ;
U$\leftarrow$ = la sélection de U est essentiellement déterminée par des contraintes provenant du niveau cible.

De plus, j'utiliserai les abréviations suivantes : s = sens, LP = lexie profonde, FL = fonction lexicale, LG = lexie grammaticale et LC = lexie collocationnelle.

[8] C'est à dessein que j'utilise le terme vague d'*unité linguistique*, plutôt que celui de *lexie*. Nous allons voir en effet qu'il n'est pas seulement question ici des choix lexicaux proprement dits.

Je présente maintenant le tableau général de la lexicalisation, étape par étape, à partir des trois phases de transition identifiées dans la section 3.1.

1) Transition conceptuelle : sélection de →s, →s← et s←

Un sens →s est introduit dans le message lorsque le module de génération profonde a déterminé qu'il était nécessaire que l'information correspondante sur l'état du monde soit communiquée. On peut supposer que tous les sens introduits dans la Figure 2 ci-dessus sont de ce type.

Un sens →s← est introduit pour satisfaire un double but. D'abord, il s'agit d'exprimer une information que le module de génération profonde a identifiée comme devant être communiquée, en fonction de contingences extralinguistiques. Ensuite, il s'agit de répondre à certains besoins purement grammaticaux. Ainsi, la spécification de la localisation temporelle d'un fait peut avoir été calculée lors de la planification si le contenu informationnel correspondant a été identifié comme important. Mais elle est aussi nécessaire, pour des raisons strictement grammaticales, si le fait en question doit être exprimé, aux niveaux subséquents de représentation, par un verbe fini.

Un sens s← est, quant à lui, uniquement introduit en fonction de contraintes grammaticales. Nous nous trouvons ici un peu dans le même cas que précédemment, mais avec une génération profonde qui n'aurait pas « éprouvé le besoin » de calculer un sens grammatical (temps grammatical, sens de la détermination définie/indéfinie, etc.) requis pour la suite de l'exécution.

Il est essentiel de noter que les sens ne sont pas de type →s,→s← ou s← de façon intrinsèque, mais qu'ils acquièrent ce statut dans le contexte d'un processus de génération donné. Il est clair, cependant, que les sens lexicaux ont tendance à fonctionner comme des sens →s ou →s←, et les sens grammaticaux, comme des sens s←.

Remarquons finalement que, dans ce modèle, aucune sélection lexicale n'est effectuée à ce niveau de transition. Ce qui est choisi ici ce sont les sens et non les lexies de la langue. Le module de génération de surface a encore, à cette étape de l'exécution, toute latitude pour sélectionner les lexies qu'il veut, pourvu que le résultat final exprime bien le message qui lui est fourni en entrée dans la représentation sémantique.

2) Transition sémantique : sélection de →LP, →LP←, LP←, →FL, →FL← et FL←

Une →LP est une lexie profonde introduite dans l'arbre syntaxique profond strictement pour exprimer une partie du message. Un tel cas est finalement assez rare. C'est peut-être le cas de certains connecteurs qui laissent peu de liberté quant au choix de la lexie à utiliser. Par exemple, le sens exprimé par *et* dans *Papa pique et maman coud* pourrait difficilement être exprimé par une autre lexie du français.

Une →LP← représente le cas le plus normal pour ce niveau de transition. C'est une lexie qui est choisie à la fois en fonction de son

contenu sémantique et en fonction de son comportement syntaxique. Par exemple, pour réaliser la racine de l'arbre syntaxique de la Figure 3 à partir de la configuration de sens 'baisser**II.4** de façon importante**I.2**' de la Figure 2, on doit choisir un verbe, CHUTER**3**, plutôt que le nom correspondant, CHUTE**I(B)5**.

Une LP←, tout en étant une lexie profonde, est choisie strictement pour des raisons configurationnelles. En français, c'est le cas de quelques lexies comme FOIS, dans *Il a éternué trois fois*. Cette lexie est un outil syntaxiquement nécessaire pour réaliser une forme de « quantification verbale », mais il n'existe aucune motivation sémantique pour son emploi (comparer avec *trois éternuements*).

Une fonction lexicale de type →FL est introduite dans l'arbre syntaxique profond du fait de la présence d'un sens collocationnel (comme l'intensification, par exemple) dans la représentation du message (c'est-à-dire, dans la représentation sémantique). C'est le cas, par exemple, de **Magn**'vitesse' dans la Figure 3.

Une →FL← répond, elle, a un double besoin : elle réalise un sens collocationnel et remplit une fonction grammaticale bien précise dans l'arbre syntaxique profond. C'est le cas d'une FL comme **Real**, qui correspond à un « verbe de réalisation » signifiant à peu près 'satisfaire à/réaliser'. Par exemple, **Real$_1$**(*mission*) = *accomplir*.

Une FL←, au contraire, ne fait que jouer un rôle syntaxique. Elle n'est aucunement sélectionnée pour exprimer un sens. C'est le cas de la fonction **Oper**, que nous avons déjà rencontrée. Ainsi, une alternative à l'utilisation de CHUTER**3** comme racine de l'arbre syntaxique profond dans mes exemples serait de sélectionner le nom correspondant, CHUTE**I(B)5**, et d'en dériver une construction verbale *verbe support + nom* au moyen de la fonction lexicale **Oper$_1$**. Cela nous donnerait, comme alternative à l'arbre de la Figure 3, l'arbre « paraphrastique » de la Figure 5 ci-dessous, qui est la représentation syntaxique profonde de la phrase (2).

(2) *Le cours du dollar a fait une chute brutale vers 13h00.*

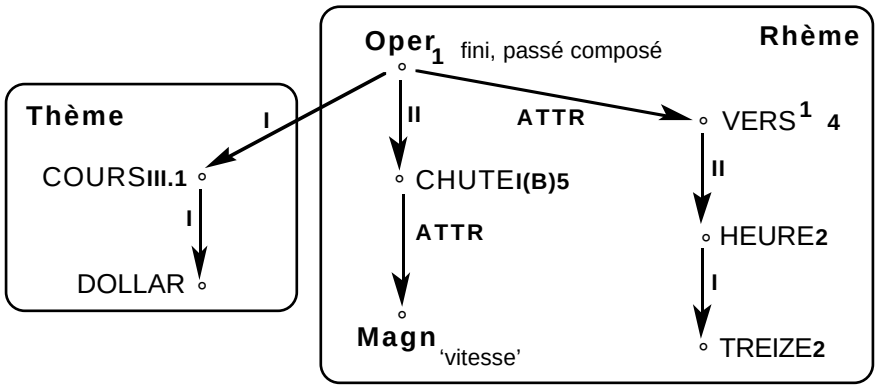

Figure 5. Représentation syntaxique profonde de (2)

La mise en correspondance, dans un contexte de génération de texte, d'arbres comme ceux des Figures 3 et 5 est un problème passionnant, dont la solution passe par le recours à des modèles lexicaux et grammaticaux de la langue très puissants. On trouvera dans (Iordanskaja *et al.* 1996) quelques ébauches de solution selon l'approche Sens-Texte, basées

notamment sur les expérimentations effectuées dans le cadre du développement des systèmes GOSSIP et LFS.

3) Transition syntaxique profonde : sélection de →LG, →LG←,LG←, →LC, →LC←et LC←

Une →LG est introduite dans la phrase uniquement afin d'exprimer une configuration syntaxique profonde. Il s'agit par exemple de l'introduction du pronom IL dans *Il a regardé Jules*.

Une →LG← est sélectionnée à la fois pour exprimer une configuration syntaxique profonde et pour répondre à des contraintes structurales imposées au niveau de l'arbre syntaxique de surface. Il s'agit par exemple du pronom clitique SE dans *Il se regarde*, qui exprime le complément du verbe et est en même temps sélectionné du fait de l'impossibilité de dire **Il regarde lui*. L'introduction de l'auxiliaire AVOIR dans les exemples (1a-c) représente un cas encore plus typique de ce type de sélection lexicale : cet auxiliaire participe à l'expression du temps grammatical et « fait tenir » la structure de l'arbre syntaxique de surface.

Le cas le plus typique de sélection d'une LG← est représenté par les prépositions vides, dont l'introduction est gouvernée par le régime syntaxique d'une lexie (*cours **du** dollar*, etc.). Une sélection de ce type se fait directement sous le contrôle de l'information encodée dans le lexique de la langue (le régime, dans la terminologie Sens-Texte, et la sous-catégorisation, dans nombre d'autres approches formelles).

Une sélection de type →LC correspond au choix d'une valeur particulière d'une fonction lexicale (***faire** une chute*, *chuter **brutalement***, etc.), choix qui est fait à partir des éléments déjà sélectionnés au niveau précédent (lexie qui est la base de la collocation et fonction lexicale elle-même).

Il peut cependant arriver qu'une valeur de FL soit aussi choisie en fonction de contraintes imposées par le niveau syntaxique de surface, ce qui nous donne une sélection de →LC←. Ainsi, les deux valeurs possibles de **Oper$_1$**(*vacances*), *être* et *passer*, se distinguent par le fait que la seconde ne peut être choisie que si *vacances* est accompagné d'un modificateur syntaxique :

(3) *Il est en vacances.* vs. *Il est en vacances à la montagne.*
?*_*Il passe des vacances.* vs. *Il passe des vacances à la montagne.*

Il existe un dernier cas : celui de la sélection de LC←. Il s'agit de valeurs de fonctions lexicales choisies strictement en fonction de considérations syntaxiques de surface, sans avoir été anticipées au niveau précédent. Dans une version antérieure du présent texte (celle présentée à TALN'98), je postulais la possibilité théorique d'un tel type de sélection, sans pouvoir toutefois en donner des exemples concrets. S. Kahane m'a fort justement fait remarquer que de tels cas étaient finalement assez courants, si l'on prend en considération le fait que certaines lexies font appel aux fonctions lexicales dans le cadre de leur régime. Par exemple, la lexie ABCÈS**I** — *abcès de X sur la partie de son corps Y* — possède le régime suivant :

X = I	Y = II
1. de N 2. A_{poss}	1. **Loc $_{in}$** N

Je fais ici usage du formalisme d'encodage des régimes (tableau de régime) utilisé en lexicographie Sens-Texte (Mel'čuk *et al.* 1995). Le tableau ci-dessus indique que le premier actant sémantique de la lexie en question peut être exprimé en surface par un groupe prépositionnel en *de* (*l'abcès* ***de Jules***) ou par un adjectif possessif (***son*** *abcès*). Ce qui doit nous intéresser ici c'est la réalisation de surface du second actant. En effet, il va se réaliser par un groupe prépositionnel contrôlé par un **Loc$_{in}$** du nom exprimant cet actant. La fonction lexicale **Loc$_{in}$** correspond à une préposition locative et sa valeur ne peut être connue qu'en fonction de la lexie particulière qui sera choisie pour exprimer ici l'actant Y (*abcès* ***sur*** *la peau*/***à***<?***dans***>*l'aine*, etc.). On a donc un cas intéressant de LC← (la valeur de **Loc$_{in}$**) qui est choisie comme le serait une préposition régie (pour réaliser au niveau syntaxique le second actant d'une lexie), mais en tant que valeur d'une fonction lexicale. Cela n'aurait en effet pas de sens de lister toutes les valeurs possibles de la préposition dans le régime de ABCÈS, avec des contraintes expliquant avec quel nom de Y va chaque préposition. Nous sommes véritablement ici en présence de phénomènes collocationnels, devant être décrits au moyen des fonctions lexicales.

CONCLUSION

L'approche de la lexicalisation que je viens de présenter incorpore dans un modèle commun et unifié tous les types de choix lexicaux impliqués en génération. En cela, elle se distingue d'autres solutions théoriques proposées où, notamment, il est postulé qu'un modèle de la lexicalisation des lexies profondes peut être développé indépendamment de la prise en considération d'autres types de lexicalisation, comme celle des lexies collocationnelles — voir par exemple, (Stede 1996:29).

Un aspect important du problème de la lexicalisation demeure en suspens : celui de la sélection de l'information devant être intégrée dans le message que la phrase va exprimer. D'où viennent les sens ? Comment sont-ils sélectionnés ? Le premier avantage que je vois dans le modèle proposé ici est justement d'identifier ce qui relève du problème de la lexicalisation comme tel et ce qui relève du problème de la construction du message linguistique. Ce dernier problème est trop complexe pour être analysé ici autrement que par la caractérisation de la transition conceptuelle qui est faite à la section 3.2. Je tiens toutefois à souligner que je suis en faveur d'une approche basée sur une « sous-spécification » du message transmis au générateur de surface. En effet, la langue nous impose, lorsque nous effectuons des choix lexicaux et syntaxiques, d'exprimer certaines informations que, sinon, nous n'aurions pas éprouvé le besoin de communiquer. Pour modéliser cet état de fait, il faut laisser au générateur de surface la possibilité d'aller rechercher de l'information complémentaire à un niveau très profond (et non linguistique) de représentation des connaissances. L'approche que je défends ici est

similaire à celle présentée de façon très claire et convaincante dans (Zock 1996).

Au niveau pratique, nous sommes en train de développer, à l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal, un générateur de surface, basé sur une grammaire et un dictionnaire Sens-Texte du français. La finalité de ce projet est double : utiliser la génération de texte comme outil de développement des modèles linguistiques Sens-Texte et implémenter le modèle procédural de la lexicalisation présenté ici afin d'en démontrer la validité théorique et pratique.

RÉFÉRENCES

CAMPBELL M. S. (1997) : ""An Enjoyable Game": How HAL Plays Chess", in D. G. Stork (ed.) *Hal's Legacy: 2001's Computer as Dream and Reality*, Cambridge, The MIT Press, pp. 75-98.

DE SMEDT K. (1990) : "IPF : An Incremental Parallel Formulator", in R. Dale, C. S. Mellish, M. Zock (eds.) *Current Research in Natural Language Generation*, London, Academic Press, pp. 167-192.

ELHADAD M. (1992) : Using Argumentation to Control Lexical Choice : A Functional Unification Implementation, Thèse de PhD, Columbia University, New York.

ELHADAD M., MCKEOWN K., ROBIN J. (1997) :"Floating Constraints in Lexical Choice", *Computational Linguistics*, 23:2, pp. 195-239.

IORDANSKAJA L., KIM M., POLGUÈRE A. (1996) : "Some Procedural Problems in the Implementation of Lexical Functions for Text Generation", in (Wanner 1996), pp. 279-297.

IORDANSKAJA L., KITTREDGE R. I., POLGUÈRE A. (1991) : "Lexical selection and paraphrase in a Meaning-Text generation model", in C. L. Paris, W. R. Swartout, W. C. Mann (eds.) *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, Dordrecht, Kluwer, pp. 293-312.

IORDANSKAJA L., POLGUÈRE A. (1988) :"Semantic Processing for Text Generation", *Proceedings of the First International Computer Science Conference*, Hong Kong, pp. 310-318.

KITTREDGE R. I., GOLDBERG E., KIM M., POLGUÈRE A. (1994) : "Sublanguage Engineering in the FoG System (Poster paper)", *Proceedings of the Fourth Conference on Applied Natural Language Processing*, Stuttgart, pp. 215-216.

KITTREDGE R. I., POLGUÈRE A. (1991) : "Dependency Grammars for Bilingual Text Generation : Inside FoG's Stratificational Models", *Proceedings of the International Conference on Current Issues in Computational Linguistics*, Penang, pp. 318-330.

MCKEOWN K. R., ROBIN J., KUKICH K. (1995) : "Generating Concise Natural Language Summaries", *Information Processing & Management*, 31:5, pp. 703-733.

MEL'ČUK I. A. (1995) : "Phrasemes in Language and Phraseology in Linguistics", in M. Everaert, E.-J. van der Linden, André Schenk, Rob Schrender (eds.) *Idioms: Structural and Psychological Perspectives*, Erlbaum, Hillsdale, pp. 167-232.

MEL'ČUK I. A. (1997) : *Vers une linguistique Sens-Texte. Leçon inaugurale*, Paris, Collège de France, Chaire internationale.

MEL'ČUK I. A., CLAS A., POLGUÈRE A. (1995) : *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve, Duculot.

MEL'ČUK I. A., POLGUÈRE A. (1987) : "A Formal Lexicon in the Meaning-Texte Theory (or How to Do Lexica with Words)", *Computational Linguistics*, 13:3-4, pp. 261-275.

NICOLOV N., MELLISH C., RICHIE G. (1997) :"Approximate Chart Generation from Non-Hierarchical Representations", in R. Mitkov, N. Nicolov (eds.) *Recent Advances in Natural Language Processing: Selected Papers from RANLP'95*, Amsterdam, Benjamins, pp. 273-294.

NOGIER J.-F., ZOCK M. (1992) : "Lexical choice by pattern matching", *Knowledge Based Systems*, 5:3, pp.200-212.

POLGUÈRE A. (1997) : "Meaning-Text Semantic Networks as a Formal Language", in (Wanner 1997b), pp. 1-24.

POLGUÈRE A. (1998) : "La théorie Sens-Texte", *Dialangue*, 8-9, Université du Québec à Chicoutimi, pp. 9-30.

STEDE M. (1995) : "Lexicalization in natural language generation: a survey", *Artificial Intelligence Review*, 8, pp. 309-336.

STEDE M. (1996) : Lexical semantics and knowledge representation in multilingual sentence generation, Thèse de PhD, University of Toronto, Toronto.

STEDE M. (1998) : "A Generative Perspective on Verb Alternations", *Computational Linguistics* (Special Issue on Natural Language Generation), 24:3, pp. 401-430.

WANNER L. ed. (1996) : *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam, Benjamins.

WANNER L. (1997a) : Exploring Lexical Resources for Text Generation in a Systemic Functional Language Model, Thèse de PhD, Universität des Saarlandes, Saarbrücken.

WANNER L. (1997b) : *Recent Trends n Meaning-Text Theory*, Amsterdam, Benjamins.

ZOCK M. (1996) : "The Power of Words in Message Planning", *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, Conpenhague, pp. 990-995.