Université de Paris VII - Denis Diderot

U.F.R de Linguistique

Générer des collocations

Jacques Steinlin

Mémoire
pour l'obtention du DEA
de Linguistique Théorique, Descriptive et Automatique
option Linguistique et Informatique

Septembre 2003

Sous la direction de Sylvain Kahane

Remerciements

Qu'il me soit permis d'exprimer ici toute ma reconnaissance ainsi que mon estime à Sylvain Kahane, avec qui il fait bon travailler.

Bon nombre de membres de TALaNa m'ont prodigué de précieux conseils sur bien des points théoriques ou pratiques : Laurence Danlos, Adil El Ghali, Laurent Roussarie, Laura Kallmeyer, Laurence Delort, Céline Raynal.

Merci également à François Lareau dont la lecture de mon manuscrit a permis de supprimer imprécisions ou âneries (celles y demeurant sont mon fait) et à François Toussenel pour son support technique.

Quoiqu'éloigné géographiquement, Alain Polguère n'a jamais manqué de répondre à mes interrogations lexicographiques et lexicologiques. Son enseignement s'est avéré précieux.

Enfin, Pascal Amsili m'avait fortement encouragé à entreprendre ce DEA. Je me félicite d'avoir suivi son conseil et le remercie pour son encadrement exceptionnel durant les deux années qui ont précédé.

Aux membres de TALaNa non cités, je souhaite témoigner du plaisir que j'ai à travailler parmi eux. En particulier, Lucie Barque, qui depuis trois ans s'est révélée être une binôme efficace à la compagnie chaleureuse.

Table des matières

Introduction **1**

1 Collocations **3**
1.1 Qu'est-ce qu'une collocation ? 3
1.1.1 Règles lexicales 3
1.1.2 Compositionnalité 4
1.2 Description des liens lexicaux 6
1.2.1 Lien lexical 6
1.2.2 Langage des fonctions lexicales 7
1.2.3 Caractérisation sémantique des fonctions lexicales 12

2 Génération **15**
2.1 Généralités 15
2.1.1 Tâches à accomplir 17
2.1.2 Architectures 18
2.2 La lexicalisation 20
2.2.1 Interface conceptuel/lexical 21
2.2.2 Interdépendance des choix lexicaux et syntaxiques 21
2.2.3 Patrons de phrases 22
2.3 G-TAG et la Théorie Sens-Texte 25
2.3.1 G-TAG 25
2.3.2 La Théorie Sens-Texte 32
2.3.3 Confrontation des deux modèles 39

3 Génération des collocations **43**
3.1 Intégration dans G-TAG 43
3.1.1 FL← : $\mathbf{Oper}_{\mathrm{i}}$ 44
3.1.2 LC← : $\mathbf{Loc}_{in}$ 45

3.1.3 →FL : **Magn** 47
3.2 Interface conceptuel/sémantique 50
3.2.1 Approche conceptuelle 50
3.2.2 Approche lexicale 51

Conclusion **59**

Bibliographie **61**

Introduction

Le travail présenté dans ce mémoire a pour objet la sémantique lexicale dans le cadre de la Génération Automatique de Textes (GAT). L'évocation de cette dernière discipline suscite généralement chez les personnes étrangères au Traitement Automatique des Langues une certaine incrédulité : pourquoi et surtout comment générer des textes ? Les questions relatives à l'analyse des langues naturelles sont en revanche souvent intuitivement mieux acceptées. D'ailleurs la recherche scientifique en linguistique se place préférentiellement dans une démarche d'analyse que de génération. Pourtant la génération offre nombre d'intérêts, tant pratiques que théoriques.

Les raisons pratiques concernent la conception de systèmes informatiques capables de générer des textes en langue naturelle. Ceux-ci représentent un enjeu industriel dans plusieurs domaines. Aussi bien les informations numériques, comme des cours de bourse ou des relevés météorologiques, que la réponse faite à une requête sur une base de données ou bien l'énumération des caractéristiques techniques d'un appareil peuvent faire l'objet d'une textualisation. L'utilisateur dispose alors des informations utiles dans un code bien plus intelligible que ne sauraient l'être des listes de nombres et mettant en avant les informations saillantes.

Les raisons théoriques tiennent au fait que la communication verbale nécessite deux participants, le locuteur et l'interlocuteur. De même que l'on s'intéresse à modéliser le travail que fait l'interlocuteur (l'analyse linguistique), il paraît légitime de tenter d'établir un modèle de la locution.

Le problème de la génération de collocations est intéressant à ces deux points de vue. D'une part, les collocations sont des idiomatismes qu'un système de génération se doit de produire afin que les textes générés soient naturels, voire simplement lisibles. Les deux énoncés suivants, par exemple, sont non idiomatiques :

Jean a donné ses excuses à Marie.
Jean a offert ses excuses à Marie.

On parvient néanmoins à comprendre leur sens, mais n'importe quel locuteur

du français les rejetera comme incorrects. Plutôt que *donner* ou *offrir*, on emploiera préférentiellement le verbe *présenter* :

Jean a présenté ses excuses à Marie.

Il est donc souhaitable qu'un système de génération soit à même d'en tenir compte. Mais la notion de collocation est également une question théorique qui fait l'objet d'une description dans la **Théorie Sens-Texte**, description articulée avec l'ensemble du modèle linguistique. Les collocations y prennent une place importante du point de vue descriptif, le collocatif n'étant pas seulement un idiomatisme qu'il faut connaître mais aussi un élément essentiel au plan sémantique.

Nous nous intéresserons donc tout d'abord aux questions de description des collocations. Après avoir défini le phénomène de cooccurrence lexicale restreinte, nous nous attacherons aux problèmes d'encodage des liens lexicaux par les **fonctions lexicales**.

Nous présenterons ensuite la Génération Automatique de Textes en nous attachant plus particulièrement à l'opération de lexicalisation. Nous confronterons notamment la Théorie Sens-Texte au formalisme **G-TAG**.

Enfin, la dernière partie sera consacrée aux possibilités d'intégration de la notion de fonction lexicale dans G-TAG. Ceci nous amènera à nous interroger sur les questions très générales de représentations conceptuelles.

Chapitre 1

Collocations

Dans ce chapitre nous allons d'abord caractériser la notion de **collocation** puis nous présenterons **les fonctions lexicales**, l'outil proposé par la Théorie Sens-Texte pour modéliser les collocations.

1.1 Qu'est-ce qu'une collocation ?

Le terme de « collocation » désigne deux phénomènes assez différents.

La première acception sert à décrire des relevés statistiques effectués sur des corpus. L'apparition fréquente dans un contexte plus ou moins étroit de deux unités lexicales données peut servir de fondement à des considérations sémantiques ou idéologiques. Dans le domaine de la statistique lexicale, « collocation » signifie donc simplement « que l'on trouve au même endroit ».

La seconde acception couvre un sous-ensemble particulier de collocations que l'on appelle aussi **cooccurrences lexicales restreintes**. Quoique vraisemblablement bien représentées dans les corpus, l'intérêt qu'on trouve à celles-ci tient davantage aux contraintes qu'elles font peser sur l'utilisation du lexique. Il s'agit d'un savoir lexical, d'une propriété du lexique. Tout locuteur natif d'une langue donnée est capable de les reconnaître.

1.1.1 Règles lexicales

De même qu'il existe des règles de bonne formation sémantiques, syntaxiques, morphologiques et phonologiques, il existe des règles régissant le lexique. Les exemples (1) à (4) illustrent des entorses faites aux règles de bonne formation respectivement phonologiques, morphologiques, syntaxiques et sémantiques.

(1.a) * la eau est fraîche
(1.b) l'eau est fraîche
(2.a) * les cheval galope
(2.b) les chevaux galopent ou c) le cheval galope
(3.a) * est amoureuse marie
(3.b) marie est amoureuse
(4.a) # d'incolores idées vertes
(4.b) [il est souvent impossible de corriger une malformation sémantique]

De semblables règles existent aussi pour le lexique. Les énoncés suivants sont à cet égard mal formés :

(5.a) Jean a #poussé/émis un son
(5.b) Jean a poussé/#émis un hurlement
(6.a) Ce souvenir reste pour lui une plaie vive/#béante
(6.b) Il restait allongé avec une large plaie/#plaie vive qui saignait
(7.a) Il régnait dans la salle un silence pesant/#malheureux
(7.b) Il lui fit alors un compliment #pesant/malheureux sur sa tenue
(8.a) Son départ lui a infligé/#inspiré une vive déception
(8.b) Cette puissance inspire/#inflige une crainte malfaisante à Pharaon

On peut constater que ces énoncés n'enfreignent ni les règles sémantiques ni les règles syntaxiques. Les énoncés bien formés et mal formés suivent tous les mêmes règles syntaxiques et restent interprétables. Ceux qui sont marqués comme étant mal formés commettent des infractions aux règles lexicales. Toutefois, à ce stade nous ne disposons pas de moyen permettant de caractériser plus précisément ces infractions.

1.1.2 Compositionnalité

Le principe de compositionnalité peut nous aider à comprendre le fonctionnement des cooccurrences lexicales restreintes.

Le principe de compositionnalité en sémantique spécifie que le calcul du sens d'un énoncé peut être réalisé en composant les sens de ses unités.

Les énoncés suivants, par exemple, ne sont pas compositionnels :

– « chemin de fer »
– « pomme de terre »
– « levée de boucliers »

Un *chemin de fer* n'est ni un chemin, ni quelque chose qui serait en fer. Il s'agit d'un phrasème, qui par le passé devait entretenir des liens avec les lexèmes qui le composent, mais qui doit être traité aujourd'hui en bloc.

Les cooccurrences lexicales restreintes sont semi-compositionnelles. Par exemple l'énoncé « long voyage » semble compositionnel si l'on s'accorde à dire que « long » signifie ici 'qui dure longtemps'. Autrement dit :

« long voyage » ≈ 'voyage qui dure longtemps'

Si l'on examine les énoncés « livre long », « longue description » et « long travail », on ne parvient pas à les paraphraser de la même façon :

« livre long » ≈ 'livre que l'on met longtemps à lire'
« longue description » ≈ 'description détaillée'
« long travail » ≈ 'travail qui nécessite un grand intervalle de temps pour être effectué'
« long détour » ≈ 'détour sur une grande distance'

Par conséquent, chaque occurrence de « long » présentée ici requiert une définition propre. Une esquisse de travail lexicographique pourrait être :

$long_1$: 'qui dure longtemps'
$long_2$: 'qui met longtemps à être lu'
$long_3$: 'qui est détaillé'
$long_4$: 'qui nécessite un grand intervalle de temps pour être effectué'
$long_5$: 'qui s'effectue sur une grande distance'

Cette approche préserve le principe de compositionnalité. Le fait que les définitions soient composées de façon *adhoc* n'est pas un problème en soi. Il donne une indication du lien qui existe entre les lexies *voyage*, *détour*, *livre*, *description* et *travail* d'une part et les différentes acceptions de « long ». On peut donc dire que le sens de « long » est une **fonction** de la lexie qu'il modifie. On peut franchir un pas supplémentaire et dire que pour chacune des acceptions de « long » la même fonction est à l'œuvre. Cela nous permettrait d'expliquer que l'on accepte bien les énoncés suivants :

« il vaut mieux faire un voyage qu'un détour »
« il vaut mieux faire un long voyage qu'un long détour »

mais pas :

? ?« il vaut mieux faire un voyage intéressant qu'un long détour »
? ?« il vaut mieux faire un long voyage qu'un détour superflu »

Le parallèle est rendu impossible dans les deux derniers énoncés par le fait qu'il n'est pas pertinent de contraster les modifieurs respectifs de *voyage* et *détour*. Or les définitions de $long_1$ et $long_5$ ne sont pas non plus contrastives et devraient donc bloquer la pertinence du parallèle[1]. Ainsi, poser que le sens de « long » est une fonction de la lexie qu'il modifie et que cette fonction est la même pour *voyage* et *détour* nous autorise à poser que $long_1$ entretient le même rapport avec *voyage* que $long_5$ vis-à-vis de *détour* et explique la pertinence du parallèle.

1.2 Description des liens lexicaux

Dans les lignes qui suivent, nous développons les notions de **lien lexicaux** et **fonction**. En particulier, nous présentons le langage des **fonctions lexicales** utilisé pour l'encodage des liens lexicaux.

1.2.1 Lien lexical

Comme nous l'avons vu dans la section 1.1.1, certains sens ne s'expriment pas librement. Ainsi, par exemple, le sens 'intense' est-il réalisé de façons très diverses dans les énoncés suivants :

– « un éloge appuyé »
– « les plus plates excuses »
– « une imagination débridée »

Partant du constat qu'il devait être néanmoins possible de rendre compte d'une certaine proximité sémantique de ces énoncés, Žolkovskij et Mel'čuk ont mis au jour en 1965 la notion de **fonction lexicale**.

Au niveau sémantique, il est possible de caractériser les énoncés ci-dessus de façon informelle par les expressions suivantes :

– « éloge appuyé » ≈ 'intense éloge'
– « plus plates excuses » ≈ 'excuses intenses'
– « imagination débridée » ≈ 'imagination intense'

Pour reprendre les exemples de la section précédente, on pourra poser :

[1] L'homophonie ne semble pas jouer un rôle. D'une part on trouve ce type d'énoncés contrastifs dans les proverbes (« un mauvais arrangement vaut mieux qu'un bon procès »), d'autre part on ne décèle pas l'intention humouristique propre au rapprochement de deux homophones (« la facture de ce plat est à l'image de la facture à régler : salée »).

– « long voyage » ≈ ‘voyage intense’
– « long détour » ≈ ‘détour intense’

Cette façon de représenter la sémantique des énoncés permet bien l'établissement d'un lien, toujours le même, entre le substantif et le modifieur. Il est donc possible de définir une fonction rendant compte de ce lien. C'est le point que nous allons maintenant développer.

1.2.2 Langage des fonctions lexicales

Les fonctions lexicales constituent un langage permettant de décrire les liens lexicaux. On les regroupe en deux catégories : la dérivation sémantique et les collocations.

1.2.2.1 Liens lexicaux décrits

La dérivation sémantique établit un lien paradigmatique avec la lexie considérée. Les liens les plus évidents sont ceux de synonymie (ou quasi-synonymie), de généricité (ou hyperonymie) et d'antonymie. Mais des liens sont également établis vers les dérivés syntaxiques comme le substantif associé à un verbe (ou inversement, le verbe associé à un substantif) l'adjectif associé à un verbe (ou à un nom), . . . etc.

On liera donc les unités lexicales suivantes à *aimer* :
→ *amour*
→ *détester*
→ *amoureux*(Adj)
→ *amoureusement*

Toujours dans le cadre de la dérivation sémantique, on considérera les noms typiques d'actants et de façon générale les lexicalisations des participants associés à l'unité lexicale décrite. Il y a parfois une méprise sur la nature de ce lien qui est considéré tantôt comme morphologique tantôt comme encyclopédique.

Or il s'agit bien d'un lien sémantique et non morphologique. En effet, il paraît naturel de relier *mangeur* à *manger*, car le lien est soutenu par la morphologie. Mais ce même lien se retrouve entre *conducteur* et *voiture* sans que la morphologie ne fournisse aucun appui. Le lien est sémantique, il caractérise le nom typique du premier actant sémantique de la lexie. Le lien encyclopédique existe

lui aussi, mais c'est la relation lexicale qui sert à le révéler, et non l'inverse.

Les liens collocationnels rendent compte des phénomènes idiomatiques. Il faut les distinguer des combinaisons libres, d'une part, et des locutions (phrasèmes) d'autre part.

Les expressions libres peuvent s'interpréter comme étant sématiquement compositionnelles :

> « lexie1 lexie2 » s'interprète 'lexie1' $\oplus$ 'lexie2'

En d'autres termes, la sémantique d'une expression libre constituée de deux lexies est égale à la composition des sens atomiques associés à chaque lexie. Ainsi, l'expression « ce robot est pratique » est-elle librement construite. Il est en effet possible d'associer à l'adjectif *pratique* la sémantique 'qui remplit bien sa fonction' et de prédire que n'importe quel artefact X pourra se combiner librement avec lui pour former l'expression « X remplit bien sa fonction ». Le *dictionnaire des cooccurences* de Jacques Beauchesne (Beauchesne, 2001) donne par exemple une liste de cooccurrents pour *instrument*. Celle-ci présente entre autres l'adjectif *utile*. De notre point de vue, il ne s'agit pas d'une relation lexicale spécifique, puisque l'on peut l'employer pour qualifier à peu près n'importe lexie nominale : « un renseignement utile », « une personne utile »..., etc.

A l'opposé, un phrasème est constitué de plusieurs vocables dont aucune acception ne permet de rendre compte de la sémantique de façon compositionnelle. C'est le cas d'expressions comme : *chemin de fer*, *levée de boucliers*, ...etc, déjà mentionnées.

Les liens collocationnels sont semi-compositionnels. Ils mettent en jeu une lexie appelée « base » et un lexie dépendante de celle-ci, appellée « collocatif ». On dira que la lexie « base » contrôle son collocatif. Dans les enoncés suivants nous avons noté en caractères gras la lexie base : « pousser un **soupir** », « un **cri** strident » ou « conduire une **voiture** ». Dans toutes ces expressions, c'est le choix d'une lexie particulière (la base) qui contraint le choix du collocatif pour exprimer un sens donné.

Nous allons maintenant nous intéresser à la description formelle des liens lexicaux.

1.2.2.2 Encodage des fonctions lexicales

Il existe plusieurs façons d'encoder les liens lexicaux. Nous présentons ici l'encodage qui est adopté dans le DEC.

Les fonctions lexicales se définissent formellement de la façon suivante :

$$\mathbf{F(B) = C}$$

où :

- **F** désigne le nom de la fonction lexicale
- **B** désigne la base (appelée aussi « mot-clef »)
- **C** désigne le collocatif

En réalité, ce n'est pas un collocatif qui est associé à la base via la fonction lexicale, mais un ensemble de collocatifs.

La sémantique de la fonction lexicale F peut en outre se définir de la façon suivante :

$$\text{Sém(FL)} = \frac{\text{L1}}{\text{L2}} = \frac{\text{L3}}{\text{L4}}$$

Autrement dit, la sémantique d'une fonction lexicale donnée peut se définir comme étant ce qui est commun entre le rapport `L1/L2` et le rapport `L3/L4`. Par exemple la fonction MAGN peut se définir comme suit :

$$\text{Sém(Magn)} = \frac{\text{grosse}}{\text{bagarre}} = \frac{\text{acharnée}}{\text{lutte}} = \frac{\text{grande}}{\text{bataille}}$$

On peut se reporter à (Mel'čuk I. et al., 1999) pour une énumération exhaustive des fonctions lexicales simples au nombre d'environ soixante. Voici toutefois, à titre d'exemple, quelques fonctions lexicales parmi les plus typiques :

Magn Sa sémantique est celle de l'intensification. Par exemple : Magn(*pluie*) = {*torrentielle*, *diluvienne*, ...}.

Real$_i$ Renvoie un verbe de réalisation associé à la base. Celui-ci doit prendre comme premier complément le mot-clef. L'indice fait référence au i$^{\text{ème}}$ actant de la base occupant la position de sujet du verbe. Par exemple : Real$_1$(*bataille*) = {*gagner*, *remporter*, ...}.

Oper$_i$ Renvoie un verbe support sémantiquement vide prenant pour premier complément la base. Ici encore, l'indice fait référence au i$^{\text{ème}}$ actant de la base occupant la position de sujet du verbe. Par exemple : Oper$_1$(*bataille*) = {*participer*}.

S$_i$ Donne le nom typique du i$^{\text{ème}}$ actant de la lexie base. Par exemple : S$_1$(*accuser*) = {*accusateur*}, S$_2$(*accuser*) = {*accusé*}, S$_3$(*accuser*) = {*faute*, *crime*, *méfait*,...}.

Ces fonctions lexicales peuvent être considérées comme des relations lexicales primaires. Il est possible de les combiner afin de décrire d'autres liens lexicaux plus sophistiqués. Les combinaisons en question donnent les fonctions lexicales complexes et les configurations de fonctions lexicales. Nous expliquons comment interpréter ces formules dans la prochaine section. Voici toutefois quelques exemples de fonctions lexicales complexes :

CausDegrad(*humeur*$_{1a}$) = {*alterer*}
AntiLabor$_{12}$(*morceau*$_{I.1}$) = {*assembler*, *reunir*, *recoller* [les x]}

et quelques exemples de configurations de fonctions lexicales :

Magn+**A**$_1$(*abandon*$_{II}$) = {*plein [d']*}
Magn+**IncepOper**$_1$(*abattement(1)*) = {*sombrer [dans ART]*}

Malgré cet outillage déjà très puissant, un certain nombre de collocations échappent à la description. Le travail lexicographique de la Théorie Sens-Texte étant empirique, il s'efforce avant toute chose de rendre compte d'une réalité lexicale pour tenter seulement ensuite d'en normaliser la description. Les énoncés suivants mettent en œuvre des cooccurences :

« une lutte serrée »
« une lutte intestine », « une lutte fratricide »

Comme il n'existe pas de fonctions permettant de les décrire, il faut en forger de façon *adhoc* :

[Telle que X et Y sont d'une force comparable](*lutte*) = {*serrée*}
[Telle que X et Y appartiennent au même groupe social](*lutte*)
= {*intestine*, *fraticide*}

Concernant les questions de standardisation, nous renvoyons le lecteur à (Polguère, à paraître) qui, en s'intéressant à la fonction **[à nouveau]**, identifie les critères permettant de définir de nouvelles fonctions lexicales standards simples.

Le langage que nous venons d'introduire brièvement est utilisé en lexicographie car il présente l'intérêt d'être (assez) facilement manipulable. Certaines informations véhiculées par les fonctions lexicales sont toutefois implicites. Nous allons maintenant nous intéresser à une définition plus formelle de celles-ci.

1.2.2.3 Encodage explicite

De façon implicite, les fonctions lexicales encodent plusieurs types d'informations. Comme il a été montré dans (Kahane & Polguère, 2001b) et (Kahane & Polguère, 2001c), il faut distinguer dans une fonction lexicale :

- sa sémantique (l'apport sémantique de la collocation à la base),
- sa structure syntaxique :
  - la relation de dépendance entre la base et le collocatif,
  - la partie du discours de la valeur,
  - la structure actancielle (à savoir, comment les actants de la base se combinent avec le collocatif).

Ces informations peuvent être représentées sur deux niveaux :

$$\begin{bmatrix} \text{contenu sémantique} \\ \text{patron syntaxique} \end{bmatrix}$$

Le langage pour décrire le contenu sémantique est constitué de :

- # : le sens du mot-clef donné en argument de la fonction lexicale,
- Caus, Magn, Real, ... : les sémantiques associées aux fonctions lexicales,
- 1, 2, 3, ... : les actants sémantique du mot-clef,
- Ω : un actant sémantique externe,
- ˆ : pour exprimer la dominance communicative,
- { et } : pour spécifier un élément sémantique du contexte ne prenant pas part à la sémantique de la fonction lexicale,
- (et) : afin de délimiter la portée des opérateurs.

Le langage pour la partie syntaxique est le suivant :

- V, A, N, Adv : les parties du discours,
- 1, 2, 3, ... : les dépendants syntaxiques de la valeur,
- # : le mot-clef,
- ˆ : la dépendance syntaxique.

Voici quelques exemples de fonctions lexicales selon cet encodage :

$$\mathbf{Oper}_i = \begin{bmatrix} \# \\ \text{V[i, \#]} \end{bmatrix} \quad \mathbf{Loc}_{in} = \begin{bmatrix} \text{Loc}_{in}\text{[\#]} \\ \text{Adv[ˆ\#]} \end{bmatrix}$$

$$\mathbf{Magn} = \begin{bmatrix} \text{\{\#\}ˆMagn} \\ \text{A[\#ˆ]} \end{bmatrix} \text{ ou } \begin{bmatrix} \text{\{\#\}ˆMagn} \\ \text{Adv[\#ˆ]} \end{bmatrix}$$

$$\mathbf{Caus} = \begin{bmatrix} \text{Caus[}\Omega\text{,\#]} \\ \text{V[}\Omega\text{,\#]} \end{bmatrix} \quad \mathbf{Real}_i = \begin{bmatrix} \text{Real[\#]} \\ \text{V[i, \#]} \end{bmatrix}$$

L'encodage explicite, outre qu'il autorise la définition formelle des opérations évoquées à la section précédente[2], permet de caractériser trois familles de fonctions lexicales : les fonctions à rôle syntaxique ($\mathbf{Oper}_i$), les fonctions à prédominance sémantique (**Magn**) et les fonctions à caractère à la fois syntaxique et sémantique ($\mathbf{Real}_i$).

1.2.3 Caractérisation sémantique des fonctions lexicales

Les fonctions lexicales constituent une abstraction pertinente permettant de modéliser le lien qui unit deux unités lexicales. Il faut bien voir cependant que cette modélisation traduit le type de relation entre deux lexèmes, mais ne rend pas compte en détail de la nature exacte de la relation entre les deux lexies. Nous entendons par là que la valeur sémantique de la composition entre la valeur renvoyée par une fonction lexicale syntagmatique et son mot-clé n'est que partiellement calculable.

1.2.3.1 Le vague

Le vague des fonctions lexicales sémantiques est un choix lexicologique délibéré. Il permet de mettre en parallèle des valeurs en leur affectant le même rôle sémantique auprès de la lexie qui les contrôle.

Ainsi, les deux énoncés « un long voyage » et « un long détour » ne peuvent se paraphraser de la même façon :

– « un long voyage » ≈ 'un déplacement qui dure longtemps'

– « un long détour » ≈ 'un déplacement sur une grande distance'

mais on pourra néanmoins caractériser le rôle sémantique du modifieur de la façon suivante :

$$\text{'Magn'} \oplus \text{'L'}$$

On a fréquemment le cas d'une pure sémantique de **Magn**. Par exemple, la formule ci-dessus caractérise parfaitement les énoncés ci-dessous :

– « un colère noire » ≈ 'une colère intense'

– « une peur bleue » ≈ 'une peur intense'

Les adjectifs *bleu* et *noir* sont de purs modifieurs d'intensité des sentiments. Ce n'est pas le cas dans les énoncés qui suivent :

1. « un combat archarné / terrible / meurtrier / sanglant / _sans merci_ »

[2] Les fonctions complexes, les configurations de fonctions lexicales et les fusions sont présentées comme étant des produits (cf. (Kahane & Polguère, 2001c) pour une présentation détaillée).

2. « un combat furieux / intense / âpre / rude / violent »

Tandis que les modifieurs de la seconde ligne semblent caractériser l'ardeur des combattants (en l'intensifiant), ceux de la première ligne semblent davantage intensifier une composante de sens liée à l'issue du combat. Cela est lié au fait, que les collocatifs ont une certaine sémantique en plus de leur capacité à intensifier.

1.2.3.2 Magn : la composante graduable

Sémantiquement, il est nécessaire de disposer d'une composante graduable à intensifier. (Mel'čuk, 1995) expose la façon dont s'opère la modification intensificative :

> APPLAUDISSEMENTS : puisque le nom *applaudissements* reçoit tout naturellement les modificateurs adjectivaux du type *nourris <frénétiques, clairsemés>*, où l'adjectif fonctionne comme un opérateur d'intensification/d'atténuation, sa définition doit comprendre une composante en mesure d'accepter cette qualification. Une définition avec une telle composante (en petites majuscules) se lira comme suit :
>
> > *Applaudissements de X à Y pour Z* = 'Battements de mains par X en signe d'approbation par X de Z de Y dont LA FORCE ET/OU LA FRÉQUENCE est/sont proportionnelle(s) au degré de cette approbation'.
>
> La présence de la composante '...la force et/ou la fréquence...' facilement intensifiable/atténuable reflète la cooccurrence de APPLAUDISSEMENTS non seulement avec les adjectifs cités mais aussi avec d'autres adjectifs du même type (*de rares applaudissements*, *quelques applaudissements*, ...) et, en plus, avec des noms comme SALVE, TONNERRE, TEMPÊTE (*d'applaudissements*). (pp. 98-99)

On peut citer également $peine_{II}$ qui de la même façon comporte une composante graduable permettant d'interpréter correctement « une peine sévère ». La méthodologie d'encodage stipule que l'on doit faire référence explicitement à la composante portant l'intensification lorsqu'il y a ambiguïté. Ces composantes apparaissent en indice. Il est fait référence également dans le *DEC IV* à des composantes standards notées en exposant (par exemple, $\mathbf{Magn}^{\text{temp}}(voyage) = long$). Enfin un indice numérique permet de spécifier l'actant sur lequel porte l'intensification (par exemple, $\mathbf{Magn}_1^{\text{quant}}(applaudissements) = nombreux$).

On remarque toutefois que l'absence d'indication de composante traduit souvent le fait que la lexie ne dispose pas en elle-même d'une composante intensifiable.

Le DiCo dessinant une hiérarchie de types (cf. (Polguère, à paraître b)), on peut tenter de trouver les composantes intensifiables en remontant au niveau du type. Ainsi, dans « une haine ardente », *ardente* s'analyse comme un intensifieur portant sur le type *sentiment*[3]. Dans l'énoncé « une haine tenace », *tenace* est décrit comme un intensifieur portant sur la composante standard *durée* (**Magn**$^{\text{temp}}$) mais on peut se demander pourquoi réserver un traitement priviligié à ce type de composantes puisqu'au fond elles sont, elles aussi, rattachées à un type, en l'occurence le type *fait*.

[3] La décomposition sémantique de *sentiment* comporte un groupement de sémantèmes sur lesquels peut porter l'intensification.

Chapitre 2

Génération

Nous introduisons dans ce chapitre la Génération Automatique de Textes. Après avoir exposé les principes généraux de cette discipline, nous accorderons une attention particulière au problème de la lexicalisation. Nous comparerons notamment deux approches de la génération : le formalisme G-TAG (Danlos, 1998) et la Théorie Sens-Texte (Polguère, 1998).

2.1 Généralités

On entend par générateur de textes, un système capable de produire des textes en langue naturelle. L'éventail des possibilités est important : le message d'erreur envoyé à l'utilisateur d'un programme peut s'assimiler à un texte généré automatiquement dans la mesure où il est une réponse à une opération effectuée sur la machine[1]. De façon peut-être plus pertinente, on peut aussi considérer que les programmes de publipostage font de la génération automatique de textes. Ces systèmes reposent sur des patrons à compléter à l'aide d'informations disponibles dans une base de données. Un autre programme, célèbre chez les utilisateurs d'**emacs**, *Eliza*, génère des réponses aux assertions de l'utilisateur. La figure (2.1) donne un exemple de dialogue que l'on peut obtenir. Ici, le principe est à nouveau celui des patrons : des patrons de phrases sont sélectionnés pour inclure des éléments des assertions précédentes.

Ces systèmes sont souvent relativement simples à mettre en œuvre et peuvent s'avérer très efficaces. Toutefois, ils comportent des limites importantes de flexibilité

[1]On objectera que le message n'est pas véritablement généré. Mais bien souvent les messages contiennent des variables (par exemple la date et l'heure ou le nom d'un fichier : "souhaitez-vous vraiment effacer le fichier nommé "mémoire.tex" ?").

```
- hello doctor !
- How do you do?  What brings you to see me?
- I am anxious.
- Is it because of problems at school that you say you are anxious?
- Yes, I might be late with my homework.
- Are you sure that yes you might be late with your homework?
```

FIG. 2.1 – Dialogue avec le psycho-thérapeute d'**emacs**

et surtout ils ne nécessitent pas de véritables connaissances linguistiques. Nous nous intéresserons ici seulement aux systèmes mettant en jeu une modélisation de la langue[2].

Un système de génération élaboré respecte l'interface suivante : l'*input* est abstrait (i.e. n'est pas du langage naturel) et l'*output* est un texte écrit dans une langue particulière (cf. figure(2.2)).

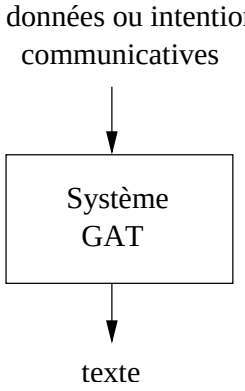

FIG. 2.2 – Principe général d'un système de génération

On voit ainsi que les systèmes à patrons ne sont pas de véritables programmes de génération puisqu'une partie de leur entrée (le patron) est déjà de nature linguistique.

L'abstraction est un objectif doublement intéressant à poursuivre. Elle permet d'une part de concevoir un système générique (c'est-à-dire un système susceptible d'assurer de la génération dans plusieurs domaines, voire de faire de la génération multilingue si l'on s'efforce de rester indépendant d'une langue particulière). D'autre

[2] En réalité, ceux-ci emploient également des patrons (des patrons syntaxiques notamment) mais à un niveau bien plus abstrait.

part, au plan scientifique elle permet de mieux mettre en évidence les processus à l'œuvre dans la production d'énoncés.

La nature exacte de l'input reste cependant assez problématique. En pratique elle est souvent fortement dépendante du type d'application que l'on souhaite produire. D'un point de vue théorique (Roussarie, 2000) identifie deux grandes orientations. Une première consiste à voir le générateur comme la simulation d'un locuteur, celui-ci ne s'exprimant que parce qu'il a une bonne raison de le faire. Ceci suppose donc une modélisation de la volonté sensée présider à la communication. La seconde consiste à concevoir un générateur comme une machine permettant de transmettre des informations.

2.1.1 Tâches à accomplir

Pour accomplir la transformation de l'entrée en un texte en langue naturelle, le système doit être capable d'accomplir un certain nombre de tâches. Celles-ci ne préjugent pas de l'architecture d'un système réel. La présentation des tâches est volontairement abstraite et tente globalement de répondre à la question : « que faut-il accomplir pour construire un message verbal à partir d'une représentation abstraite ? ». Les lignes qui suivent reprennent les points développés dans (Roussarie, 2000, pp 8-14).

Sélection du contenu profond. Il n'est pas évident que cette tâche relève de la linguistique à proprement parler. C'est d'ailleurs là-dessus que s'opposent les différentes options présentées ci-dessus concernant la nature de l'entrée. Si l'*input* est constitué d'informations, on peut considérer qu'une partie du travail est déjà accomplie. Toutefois, il est vraisemblable qu'une partie de la détermination de contenu reste à faire. Par exemple, il est nécessaire de représenter les données sous la forme de concepts. Quoiqu'il en soit, c'est une opération qui doit être effectuée tout à fait au début du traitement.

Structuration rhétorique. Comme on ne souhaite généralement pas générer des phrases isolées, il convient de prévoir un plan de document. Cette structuration est donc nécessaire pour la détermination de la segmentation en phrases ainsi que pour le choix de certaines constructions syntaxiques à l'intérieur des phrases déterminant par exemple la partition thème/rhème.

Planification syntaxique. Une étape de la génération doit consister dans le choix des arbres syntaxiques. Certaines constructions (comme les constructions passive ou clivée avec extraction) peuvent être exigées par la structuration rhétori-

que. Toutefois, les choix syntaxiques ne semblent pas devoir être traités indépendamment des choix lexicaux. En effet, le lexique guide les constructions syntaxiques (sous-catégorisation) et, inversement, les choix syntaxiques ont un impact sur les possibilités de lexicalisation.

Lexicalisation. Il s'agit ici de traduire lexicalement les concepts à exprimer. Les lexèmes contraignent de fait les structures syntaxiques et vice-versa, ne serait-ce que par leur partie du discours, comme nous venons de le dire. Ainsi le concept LOVE pourra être lexicalisé par le verbe français *aimer* ou le nom *amour*.

Ajustements morphologiques. La lexicalisation et la planification syntaxique donnent un arbre dont les feuilles, issues d'un dictionnaire, doivent être fléchies. Il faut également procéder aux agglutinements (*de+le*) ou à l'élision de certaines voyelles (*l'*, *d'*, ...).

Aggrégation. Il s'agit d'une opération visant à éviter les répétitions. (Reiter & Dale, 1999) donne l'exemple de l'énumération de jours de pluie : « Il a beaucoup plu le 27, il a beaucoup plu le 28 et il a beaucoup plu le 29 », qui peut être aggrégé en « Il a beaucoup plu le 27, le 28 et le 29 » ou en « Il a beaucoup plu du 27 au 29 ». La création des pronoms anaphoriques est une solution pour certains problèmes d'aggrégation : « Jean a décidé que Jean partirait »

On a vu qu'il existait une certaine interdépendance entre ces différentes tâches. Il convient donc d'exposer maintenant les moyens de les organiser.

2.1.2 Architectures

(Danlos & Roussarie, 2000) rappelle que le processus global de la génération peut être divisé en deux modules :

1. la détermination du contenu à transmettre (QUOI-DIRE) ;
2. la construction du discours correspondant au dit contenu (COMMENT-LE-DIRE).

Le module QUOI-DIRE détermine le message à transmettre et produit une forme logique. Le module COMMENT-LE-DIRE détermine une formulation linguistique à partir de la forme logique. Au plan pratique, cette division coïncide avec la répartition des ressources. En effet, le premier module est fortement dépendant du domaine d'application couvert par le générateur et s'appuie notamment sur des connaissances encyclopédiques. Le second fait appel aux ressources linguistiques de la langue dans

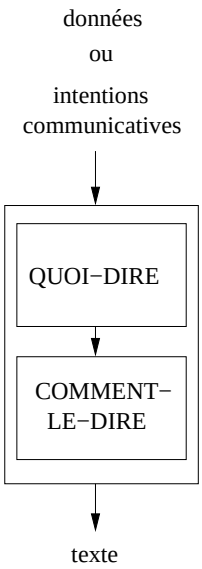

FIG. 2.3 – Principe étendu d'un système de génération

laquelle on souhaite produire le texte. De ce fait, on peut envisager la création de modules COMMENT-LE-DIRE pour la génération de textes en plusieurs langues, ceux-ci s'appuyant sur le même module QUOI-DIRE.

(Reiter & Dale, 1999) propose une répartition modulaire tri-partite des tâches énumérées dans la section précédente (cf. figure (2.4)). Cette architecture est consensuelle dans la mesure où, comme le souligne (Danlos & Roussarie, 2000), elle correspond à l'architecture des systèmes existants.

Elle ne remet pas en question la structuration bi-partite de la figure (2.3), les modules QUOI-DIRE et COMMENT-LE-DIRE correspondant respectivement à la génération profonde et à la génération de surface. Elle permet toutefois de mieux mesurer l'interdépendance entre les différents niveaux de représentation, interdépendance liée au partage des ressources. Ainsi, la question de l'aggrégation peut aussi bien avoir à être traitée au niveau conceptuel (génération profonde) par une procédure de regroupement des concepts, qu'au niveau syntaxique (génération de surface) à l'aide de constructions spécifiques (par exemple, *du ... au ...*, pour les expressions de périodes). De même, (Roussarie, 2000) indique que la structuration rhétorique peut relever du QUOI-DIRE dans la mesure où il s'agit de construire un flux d'informations pertinentes (regroupement en chunks (Reiter & Dale, 1999)) exprimées dans une structure linguistique particulière relevant, elle, du COMMENT-

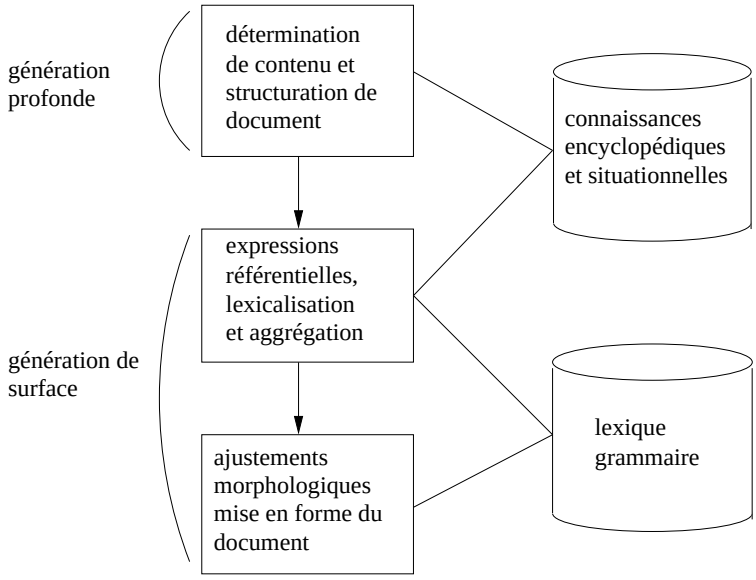

FIG. 2.4 – Architecture tri-partite

LE-DIRE.

Alors que, comme rappelé dans (Danlos & Roussarie, 2000), cela est dommageable au plan théorique, les modules sont organisés en pipeline pour des questions d'ingéniérie, interdisant de ce fait les interactions entre les composants profonds et de surface. On peut néanmoins en profiter pour se concentrer sur l'un ou l'autre des modules en postulant une certaine structure de données en entrée du COMMENT-LE-DIRE. On procèdera ainsi à rebours : le travail théorique sur le composant linguistique spécifie les informations requises au niveau précédent.

Nous considérerons pour notre part que l'entrée du COMMENT-LE-DIRE est une forme logique. En principe, celle-ci devrait être organisée rhétoriquement mais nous laisserons de côté ce point qui n'a pas d'influence sur notre propos.

2.2 La lexicalisation

Cette section est consacrée à la procédure de lexicalisation. Nous y exposons un certain nombre de questions théoriques associées à cette procédure. En particulier se posent les questions d'interface entre conceptuel et lexical ainsi que l'interdépendance des choix lexicaux et syntaxiques. Nous examinons pour finir une procédure de lexicalisation basée sur l'utilisation de patrons.

2.2.1 Interface conceptuel/lexical

Pour citer (Polguère, 1998), la lexicalisation est « le processus sélectionnant les unités lexicales ». Cette remarque, juste en soi, masque quelque peu la complexité du travail à accomplir.

Tout d'abord se pose la question de savoir ce qui constitue le point d'appui de cette opération. On pourrait penser que la lexicalisation consiste seulement à introduire des unités lexicales à la place des concepts manipulés dans les modules précédents. (Polguère, 1998) montre que ce n'est pas la seule opération requise, que l'introduction d'unités lexicales intervient en fait à différentes étapes du processus de construction de l'énoncé. Mais déjà en elle-même, l'association des concepts aux unités lexicales n'est pas triviale. En effet, les concepts ne sont nécessairement pas dans une relation bi-univoque avec les unités lexicales.

Etablir une relation de un-à-un entre les concepts et les lexèmes reviendrait à laisser de côté la synonymie, une caractéristique pourtant importante des langues. Comme le remarque (Stede, 1996), pour conserver malgré tout la richesse lexicale (la relation de synonymie entre unités lexicales) il faudrait créer autant de concepts qu'il y a d'unités lexicales, ce qui reviendrait à déplacer la question du choix lexical en amont, au moment de la construction de la représentation conceptuelle.

Par ailleurs, une correspondance bi-univoque entre concept et lexie rendrait impossible la génération multilingue évoquée dans la section précédente. Il est rare en effet de trouver entre deux lexies appartenant à deux langues un recouvrement exact de la dénotation conceptuelle.

(Stede, 1996) énumère cinq éventualités de correspondance concept/lexème (cf. figure (2.5)).

(2.5a) correspond à la relation un-un (un concept se lexicalise en un lexème). (2.5b) met en relation un concept avec plusieurs lexies entretenant des relations de synonymie ou de quasi-synonymie. (2.5c) propose de lexicaliser un concept par un lexème L1 ou bien par un hyperonyme L2 de L1. (2.5d) associe une configuration de concepts à une lexie. Enfin, (2.5e) montre un trou lexical. Cette situation nous ramène à la nécessité d'inclure le concept dans une configuration.

2.2.2 Interdépendance des choix lexicaux et syntaxiques

Un des problèmes de la lexicalisation, déjà évoqué plus haut, est qu'il existe une interdépendance entre les choix lexicaux et syntaxiques.

Une procédure de lexicalisation traitant dans un après-coup les questions syn-

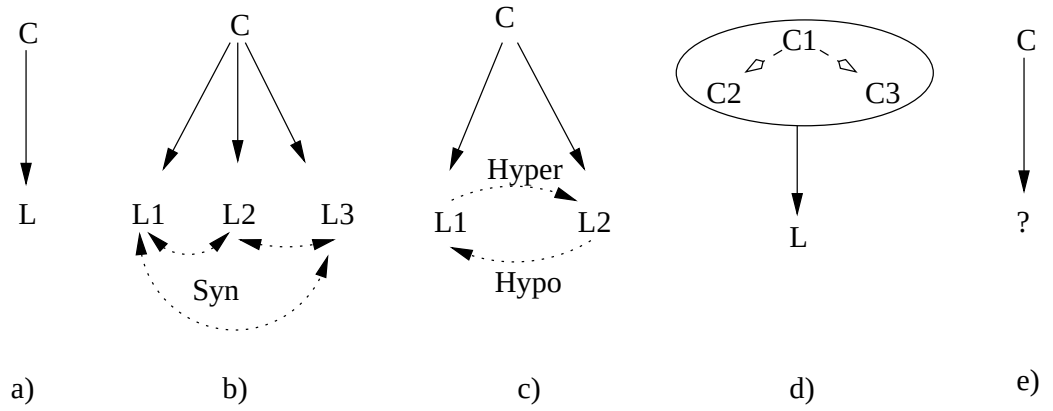

FIG. 2.5 – Interface conceptuel/lexical

taxiques se retrouvera confrontée au problème du choix des parties du discours. En effet, ceux-ci sont réalisés simultanément aux choix lexicaux. On sera alors confronté à des problèmes de construction de l'arbre syntaxique si les dépendances syntaxiques ne peuvent se caler sur les dépendances conceptuelles. À l'inverse, si l'on construisait un arbre syntaxique avant de diposer des lexies, on n'aurait pas l'assurance de pouvoir remplir toutes les positions vacantes. On peut citer également le fait qu'en génération on souhaite généralement construire des phrases dotées d'un groupe verbal, ce qui implique une lexicalisation tenant compte des parties du discours.

Par ailleurs, comme (Polguère, 1998) le montre bien, l'apparition de certains lexèmes est régie par certains autres : il nous suffit ici de mentionner le cas des prépositions sous-catégorisées par certains lexèmes (« l'amour **de** Jean **envers** Marie »).

2.2.3 Patrons de phrases

(Reiter & Dale, 2000) mentionne l'utilisation de patrons de phrases pour la génération de bulletins météorologiques.

Les patrons de phrases se présentent sous la forme de structures de traits sous-spécifiées et se combinant par unification avec d'autres structures de traits. Ainsi, par exemple, un message conceptuel donné (cf. figure (2.6)) sélectionne un patron de phrase abstrait (cf. figure (2.7)) et s'unifie avec celle-ci afin de produire une spécification de phrase (cf. figure (2.8)).

Le message apporte les informations qui doivent être instanciées dans le schéma de phrase.

$$\begin{bmatrix} \mathsf{type} : RainEventMessage \\ \mathsf{period} : \begin{bmatrix} \mathsf{month} : 05 \mathsf{year} : 1996 \end{bmatrix} \\ \mathsf{day} : \begin{bmatrix} \mathsf{day} : 27 \\ \mathsf{month} : 05 \\ \mathsf{year} : 1996 \end{bmatrix} \\ \mathsf{rainType} : heavy \end{bmatrix}$$

FIG. 2.6 – Message conceptuel (Reiter & Dale, 2000)

$$\begin{bmatrix} \mathsf{type} : PPSAbstractSyntax \\ \mathsf{head} : |fall| \\ \mathsf{features} : \begin{bmatrix} \mathsf{tense} : past \end{bmatrix} \\ \mathsf{subject} : \begin{bmatrix} \mathsf{type} : PPSAbstractSyntax \\ \mathsf{head} : |rain| \\ \mathsf{features} : \begin{bmatrix} \mathsf{definite} : |rain| \end{bmatrix} \\ \mathsf{modifier} : lexicalise(rainType) \end{bmatrix} \\ \mathsf{modifier} : \begin{bmatrix} \mathsf{type} : PPSAbstractSyntax \\ \mathsf{preposition} : |on| \\ \mathsf{object} : \begin{bmatrix} \mathsf{type} : ReferringNP \\ \mathsf{object} : day \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

FIG. 2.7 – Patron de phrase (Reiter & Dale, 2000)

Le patron de phrase comporte des informations syntaxiques (fonctions syntaxiques), lexicales (noms, verbes, prépositions régies...) et des sous-structures invoquant récursivement la procédure de lexicalisation (le trait `modifier` appelle la fonction `lexicalise` attendant un argument de type `rainType`).

Les auteurs notent que ce mécanisme ne peut valoir que lorsque le domaine d'application est suffisamment restreint. On voit en effet que le patron de phrase contient par avance beaucoup de lexèmes : *fall*, *rain* et *on*. Le message lui-même contient une information lexicale : *heavy*. Quoiqu'il en soit, le fait d'avoir prévu une position de modifieur dans le patron de phrase n'autorise pas la génération d'une unité lexicale incluant à la fois le sens de *rain* et de l'intensifieur *heavy* comme *downpour*.

$$\left[\begin{array}{ll}
\mathsf{type} : PPSAbstractSyntax & \\
\mathsf{head} : |fall| & \\
\mathsf{features} : \left[\, \mathsf{tense} : past \,\right] & \\
\mathsf{subject} : & \left[\begin{array}{l} \mathsf{type} : PPSAbstractSyntax \\ \mathsf{head} : |rain| \\ \mathsf{features} : \left[\, \mathsf{definite} : |rain| \,\right] \\ \mathsf{modifier} : |heavy| \end{array}\right] \\
\mathsf{modifier} : & \left[\begin{array}{ll} \mathsf{type} : PPSAbstractSyntax & \\ \mathsf{preposition} : |on| & \\ \mathsf{object} : & \left[\begin{array}{ll} \mathsf{type} : ReferringNP & \\ \mathsf{object} : & \left[\begin{array}{l} \mathsf{day} : 27 \\ \mathsf{month} : 05 \\ \mathsf{year} : 1996 \end{array}\right] \end{array}\right] \end{array}\right]
\end{array}\right]$$

FIG. 2.8 – Spécification de phrase (Reiter & Dale, 2000)

On peut se satisfaire d'une telle procédure pour la génération de bulletins météorologiques puisque le domaine est suffisamment réduit. Mais si l'on veut élaborer un système générique, celui-ci ne pourra pas manipuler directement des messages complexes en leur associant un patron de phrase prototypique.

La réalisation d'une interface davantage compositionnelle complexifie considérablement la tâche de la lexicalisation à tous les niveaux. (Reiter & Dale, 2000) donne l'exemple d'une lexicalisation opérée à partir d'une formule atomisée conceptuellement (cf. figure (2.9)).

Les auteurs ajoutent que la réalisation dans ce cas serait bien améliorée si l'on tenait compte de la possibilité de lexicaliser des configurations de concepts (pour

precipitationEvent(e) ∧ precipitationSubstance(e,s) ∧ material(s,water) ∧ state(s,liquid) ∧ precipitationIntensity(e, heavy) ∧ eventTime(e, time1) ∧ timeIsDay(time1, 27/5/96)

Heavy precipitation of liquid water occurred on the day 27/5/96

FIG. 2.9 – Lexicalisation atomique (Reiter & Dale, 2000)

obtenir *rain*) ou en assurant la lexicalisation des collocations (pour obtenir *fall* à la place de *occurred*). C'est la piste que nous allons suivre maintenant en nous intéressant au formalisme G-TAG ainsi qu'à la Théorie Sens-Texte.

2.3 G-TAG et la Théorie Sens-Texte

Nous allons nous intéresser ici au formalisme G-TAG (Danlos, 1998) ainsi qu'à la théorie Sens-Texte (cf. (Mel'čuk, 1997) notamment). Il s'agit de deux approches de la génération, assez dissemblables par leur nature et leurs objectifs, mais dont nous souhaitons comparer les procédures. G-TAG est un formalisme pour l'implantation d'un système de génération, alors que la théorie Sens-Texte est une théorie linguistique laissant de côté les questions procédurales pour les système informatisés, mais accordant une place importante aux collocations.

2.3.1 G-TAG

G-TAG est donc un formalisme pour la génération automatique de textes. Il a été implanté dans le générateur FLAUBERT (cf. (Meunier, 1997)). Les sections suivantes en décrivent le fonctionnement.

2.3.1.1 Couverture de G-TAG

Le formalisme couvre l'ensemble des opérations d'un module COMMENT-LE-DIRE. Il prend donc en entrée une forme logique basée sur le formalisme **Login** (cf. (Aït-Kaci & Nasr, 1986)) dont la figure (2.10) donne un exemple. C'est le choix qui a été fait pour l'implantation, néanmoins on peut considérer celui-ci comme une hypothèse, G-TAG n'y étant pas indissolublement lié.

Les concepts sont de deux types :

```
E1=: MANGER [MANGEUR => H1, MANGÉ => A1]
H1=: HUMAIN [NOM => ‘‘Jean’’, SEXE => ‘‘masc’’]
A1=: POMME []
```

FIG. 2.10 – Le formalisme **Login**

- Le type `THING` comme par exemple `HUMAIN[NOM=>String, SEXE=>String]` dont les attributs doivent être instanciés par des chaînes de caractères.
- Le type `RELATION` qui est lui-même subdivisé en `1st-order-RELATION` et en `2nd-order-RELATION`. `MANGER[MANGEUR=>HUMAN, MANGÉ=>ALIMENT]` est de type `1st-order-RELATION` et prend deux arguments dont les types sont spécifiés (`HUMAN` et `ALIMENT`).

La forme logique est traitée par le processus comme indiqué sur la figure (2.11) pour finalement produire un texte. Nous expliciterons les étapes du processus au fur et à mesure.

2.3.1.2 TAG

G-TAG se fonde sur le formalisme syntaxique TAG (cf. (Abeillé, 1993)) que nous allons brièvement présenter ici. TAG propose de représenter la phrase comme une combinaison d'arbres élémentaires dont la figure (2.12) donne quelques exemples. Chaque arbre a au moins une feuille lexicale. On distingue les arbres initiaux des arbres auxiliaires ((2.12d) est un arbre auxiliaire).

Les arbres élémentaires sont combinés selon deux opérations : la substitution pour les arbres initiaux et l'adjonction pour les arbres auxiliaires. Ces deux opérations sont représentées sur la figure (2.13).

Le nœud N1 de l'arbre (2.12c) est remplacé par l'arbre (2.12a). L'arbre (2.12d) est adjoint à l'arbre (2.12b) pour former un arbre, lui-même venant se substituer au nœud N2 de l'arbre (2.12c). Le resultat de ces opérations donne l'analyse de la phrase *René déguste une cerise* représentée sur la figure (2.14a).

On peut noter que l'ensemble des opérations ayant permis de dériver l'arbre (2.14a) ne figurent pas dans celui-ci. (2.14a) donne la structure de la phrase mais n'explique pas comment cette analyse est obtenue. L'arbre de dérivation (cf. figure (2.14b)) est la représentation annexe conservant l'historique de la dérivation. Les nœuds de cet arbre sont étiquetés par les arbres élémentaires employés lors de l'ana-

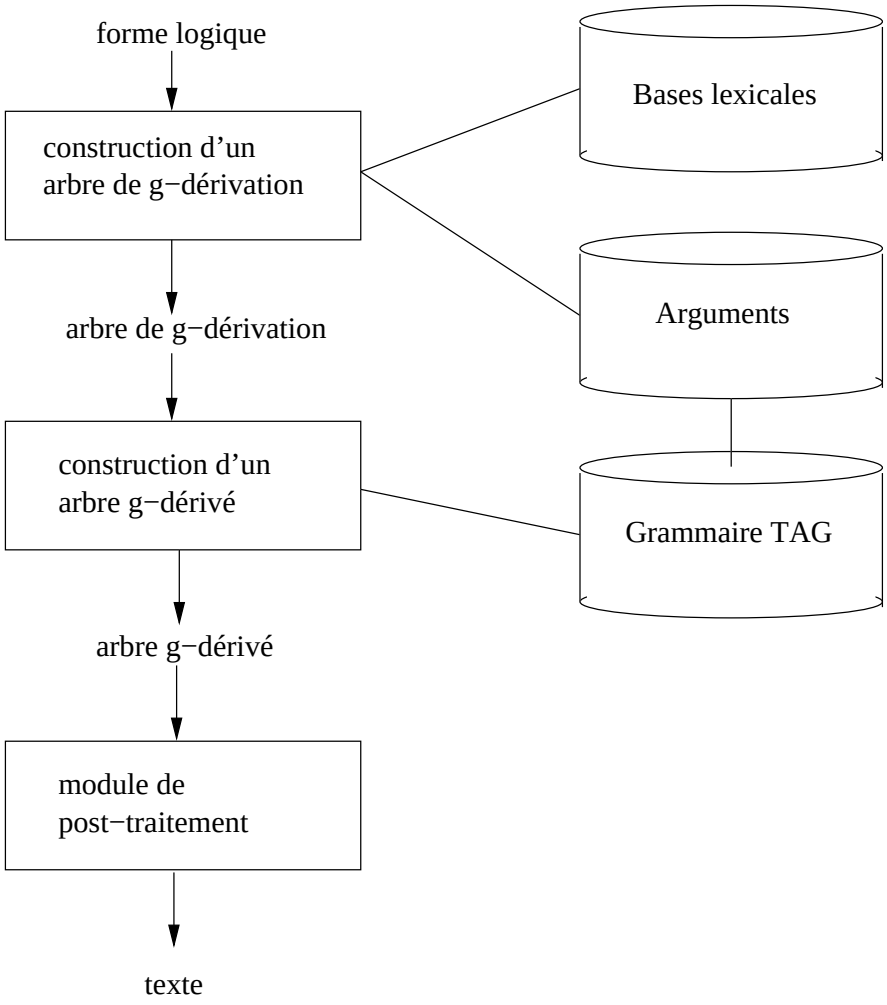

FIG. 2.11 – Principe général de G-TAG (Danlos, 1998)

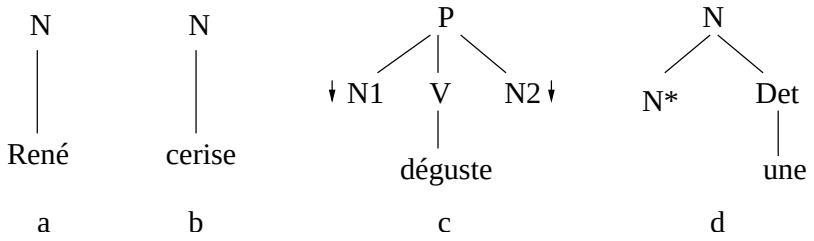

FIG. 2.12 – Arbres élémentaires TAG

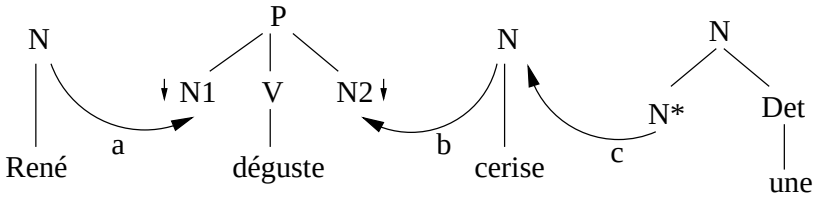

FIG. 2.13 – Opérations de substitution et d'adjonction

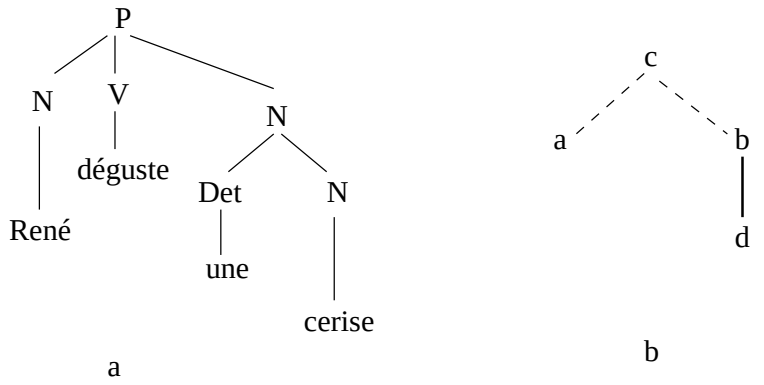

FIG. 2.14 – Arbres dérivés et de dérivation

lyse. Les traits en pointillés représentent les opérations de substitution, tandis que les traits pleins représentent les adjonctions. Cette représentation est intéressante pour la sémantique. En effet, les arbres élémentaires obéissent à quatre principes de bonne formation énumérés dans (Abeillé, 1993) :

1. Tout arbre élémentaire a au moins une tête lexicale non vide.
2. Tout prédicat contient dans sa structure élémentaire un nœud pour chaque argument qu'il sous-catégorise.
3. Tout arbre élémentaire a un correspondant sémantique non vide.
4. Un arbre élémentaire correspond à une seule unité sémantique.

En raison des principes 2 à 4, (Schieber & Schabes, 1994) propose de voir dans l'arbre de dérivation une représentation proche de la sémantique. (Danlos, 1998) rappelle que cette représentation sémantique souffre de limitations, celles-ci étant toutefois peu gênantes au plan pratique pour la génération automatique.

2.3.1.3 Arbres de G-dérivation et arbres G-dérivés

C'est la propriété des arbres de dérivation d'être une représentation proche de la sémantique qui intéresse G-TAG. Le formalisme s'inspire des arbres de dérivation pour définir des arbres de G-dérivation. La figure (2.15) donne un exemple d'arbre de G-dérivation.

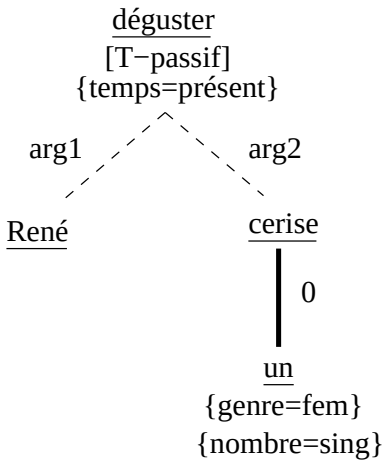

FIG. 2.15 – Arbre de G-dérivation pour « une cerise est dégustée par René »

Par rapport à la figure (2.14b) qui donnait un arbre de dérivation, on peut constater un certain nombre de changements :

1. Les nœuds sont étiquetés par des entrées lexicales et sont accompagnés de traits morphologiques et syntaxiques.
2. Les arcs spécifient les rôles thématiques.

Une fois réalisé, l'arbre de G-dérivation est transmis au module suivant qui calcule l'arbre G-dérivé (cf. (2.16)).

Un dernier module prend en entrée l'arbre G-dérivé afin de produire le texte final.

2.3.1.4 Interface conceptuel / sémantique

L'interface entre la représentation conceptuelle et la représentation sémantique (l'arbre de G-dérivation) est assurée de la façon suivante :

– Chaque concept est associé à une base lexicale contenant un ensemble d'arbres de G-dérivation sous-spécifiés.

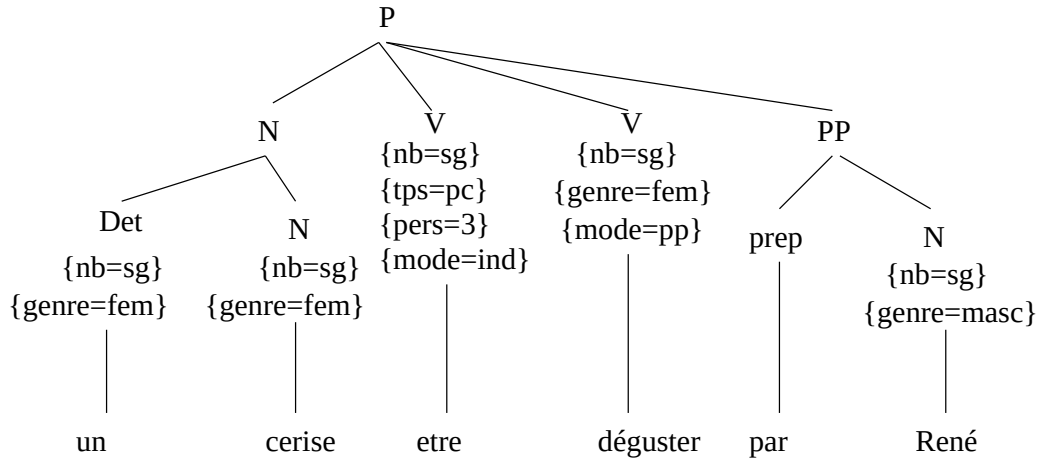

FIG. 2.16 – Arbre G-dérivé pour *une cerise est dégustée par René*

- Les arbres de G-dérivation sont identifiés par une entrée lexicale et portent des **traits forme** indiquant le type de construction qu'ils peuvent générer (texte, phrase ou groupe nominal).
- Les entrées lexicales pointent sur une famille d'arbres élémentaires.
- Une base de connaissance construite à partir de la grammaire TAG permet de connaître les fonctions syntaxiques des arguments des arbres de G-dérivation.
- Un algorithme assure la construction de l'arbre de G-dérivation par parcours de la forme logique.

Nous allons maintenant illustrer le formalisme en explicitant l'ensemble des procédures.

2.3.1.5 Un exemple

Nous souhaitons générer le texte correspondant à la représentation conceptuelle suivante donnée en figure (2.17).

```
E1=: LOVE [LOVER => H1, LOVÉ => H2]
H1=: HUMAIN [NOM => ‘‘Jean’’, SEXE => ‘‘masc’’]
H2=: HUMAIN [NOM => ‘‘Marie’’, SEXE => ‘‘fem’’]
```

FIG. 2.17 – Forme logique en entrée du générateur

Chaque instance de concept est identifiée (E1, H1 et H2). Ceci permet d'indiquer

que le premier argument de LOVE est H1, soit l'instance de concept HUMAIN, HUMAIN [NOM ⇒'Jean', SEXE ⇒ 'masc'].

Comme annoncé précédemment, chaque concept est associé à une base lexicale (**BL**) contenant un certain nombre d'arbres de G-dérivation sous-spécifiés. La figure (2.18) donne la base lexicale associée à LOVE.

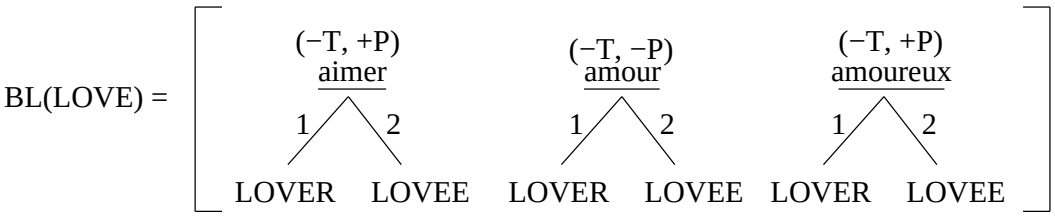

FIG. 2.18 – Base lexicale de LOVE (**BL**(LOVE))

L'algorithme de lexicalisation est appliqué récursivement en réservant une priorité aux concepts les plus élevés (dans l'ordre, 2ND-ORDER-RELATION, 1ST-ORDER-RELATION puis THING) en cas d'alternative.

Le choix est possible entre les arbres aimer et amoureux qui autorisent l'un et l'autre la génération d'une phrase (trait-forme (-T, +P)). Des heuristiques (en général stylistiques) permettent de réaliser les choix quand ils sont nécessaires[3]. Nous choisissons pour la suite l'arbre aimer. Les feuilles de cet arbre sont occupées par des nœuds variables correspondant aux positions argumentales du concept. Ceux-ci sont ensuite remplacés par les instances de concepts correspondant aux arguments de l'instance de concept LOVE, soit `H1` et `H2`. `H1` et `H2` sont alors à leur tour lexicalisés, ce qui nous donne l'arbre de G-dérivation de la figure (2.19)[4].

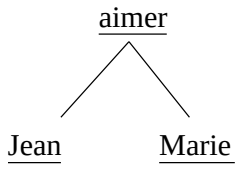

FIG. 2.19 – Arbre de G-dérivation après application de l'algorithme de lexicalisation.

Notons que si, pour des raisons stylistiques, il était requis que *Marie* soit le sujet syntaxique du verbe, l'algorithme ajouterait le T-trait `[T-passif]` à l'arbre

[3] On en trouvera une illustration dans (Danlos, 1998).

[4] Les questions grammaticales ayant trait au temps sont négligées ici.

de G-dérivation aimer. Cette possibilité nécessite que les fonctions syntaxiques des arguments des arbres dérivés soient connues au moment de la construction de l'arbre de G-dérivation, ce qui est prévu par le formalisme (cf. figure (2.11)).

Les arbres élémentaires nécessaires pour la construction de l'arbre dérivé sont disponibles dans la grammaire TAG. La figure (2.20) donne la famille d'arbres élémentaires sur laquelle pointe l'entrée lexicale aimer.

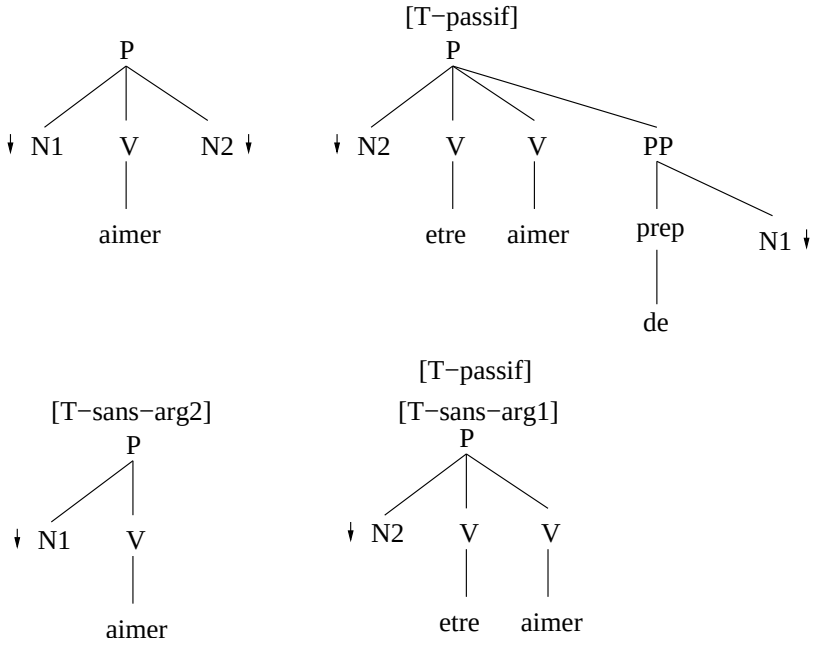

FIG. 2.20 – Famille d'arbres élémentaires associée à aimer

Ceux-ci sont combinés en suivant le schéma spécifié par l'arbre de G-dérivation. L'arbre dérivé est ensuite transmis au module de post-traitement pour divers ajustements (traitement de la morphologie, création de pronoms anaphoriques, ... etc.).

2.3.2 La Théorie Sens-Texte

Les lignes qui suivent sont une brève introduction à la Théorie Sens-Texte. Les articles et ouvrages suivants offrirons au lecteur une vision plus détaillée du modèle : (Kahane, 2001a), (Mel'čuk, 1993), (Mel'čuk, 1997), (Polguère, 1998a).

L'objet de la Théorie Sens-Texte est de théoriser la création de modèles sens-texte de langues. Ce qui la rend intéressante pour la GAT, c'est son orientation du Sens vers le Texte. Elle propose une représentation de l'activité du locuteur d'une langue donnée, celui-ci étant un générateur naturel de textes. Un des concepts majeurs

qui sous-tend cette théorie est celui de paraphrase : à un sens donné correspond plusieurs textes le réalisant en langue, ces textes étant entre eux des paraphrases synonymiques.

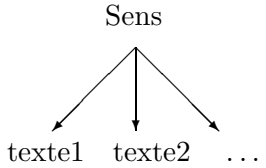

FIG. 2.21 – La paraphrase, principe fondamental de la TST

Les Textes sont les observables de la langue L (énoncés écrits ou parlés).

Le Sens (abstrait) est défini comme étant ce qui est commun à l'ensemble des Textes ayant le même sens empiriquement (cf. figure (2.21)).

L'orientation Sens → Textes permet de mettre en valeur une des propriétés des langues naturelles : la synonymie. La représentation reste toutefois réversible (texte → sens).

La correspondance du Sens au Texte se fait via plusieurs représentations intermédiaires (cf. figure 2.22).

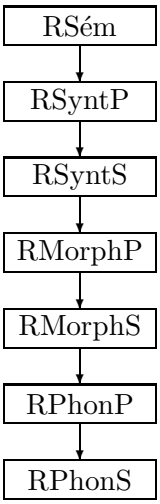

FIG. 2.22 – Niveaux de représentation

Chaque niveau encode les informations qui lui sont propres : par exemple, les représentations syntaxiques (RSyntP et RSyntS) sont des arbres exprimant les dépendances entre mots tandis que la représentation morphologique (RMorph) est une suite, ordonnée linéairement, de mots accompagnés de leur marque de flexion. On peut remarquer que le point de départ est une représentation du sens. Nous allons maintenant regarder de plus près les représentations sémantiques et syntaxiques ainsi que les règles de correspondance d'un niveau à l'autre.

2.3.2.1 Les représentations sémantique et syntaxiques

La représentation sémantique est une représentation formelle du sens. Elle se définit de la façon suivante :

La structure sémantique est un graphe acyclique, orienté et connexe.

Les nœuds du graphe sont étiquetés par des sémantèmes. Les sémantèmes sont des sens lexicaux ou grammaticaux. Ils diffèrent des concepts en ce qu'ils ne sont justement pas indépendants de la langue.

Tous les actants d'un sémantème doivent être représentés. Cela signifie qu'un sémantème ayant n actants aura n arcs partant de lui. Les arcs portent des numéros permettant de les distinguer les uns des autres.

L'exemple (2.23) est la représentation initiale des phrases :

« Jean adore Marie »
« l'amour passionné de Jean pour Marie »
« Marie est follement aimée de Jean »
« Jean est fou amoureux de Marie »
...

La figure (2.23) ne donne pas une RSem complète, il y manque en particulier la structure communicative. Il s'agit ici seulement de la structure sémantique. Il est important de bien voir que les nœuds du réseau sont des sémantèmes (i.e. des sens de la langue) et non des lexies. Ainsi le sémantème 'aimer' peut-il se lexicaliser en *aimer* ou *amour*. De même le sous-réseau 's'appeler'('personne', 'jean') peut-il être traduit par des expressions comme « la personne s'appelant Jean », « la personne qui s'appelle Jean » ou tout simplement par « Jean ». D'une façon générale, il est possible de regrouper des sémantèmes et de les représenter par une lexie ou encore un nœud peut se projeter syntaxiquement en grammème (c'est le cas de l'expression du temps).

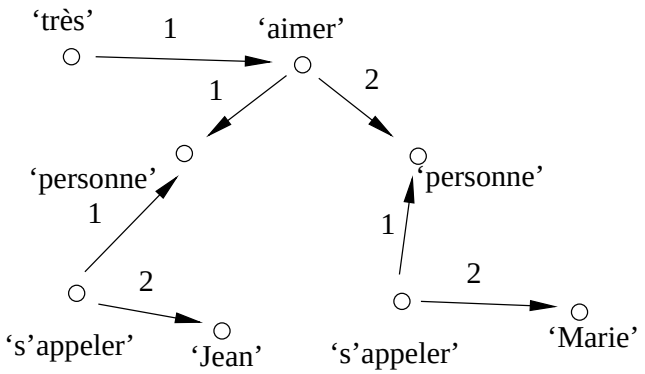

FIG. 2.23 – Structure sémantique

On synthétise à partir de la représentation sémantique la structure syntaxique profonde. Elle est définie comme suit :

La structure porteuse est un arbre de dépendance reliant non pas des constituants mais des lexies.

Les nœuds de l'arbre sont des lexies profondes ou des fonctions lexicales.

Les arcs représentent les relations actancielles profondes universelles.

Les nœuds représentent seulement des items lexicaux pleins ; les prépositions régies ou les collocations n'apparaîssent pas dans les représentation. La figure (2.24) donne des projections possibles du réseau (2.23) en syntaxe profonde.

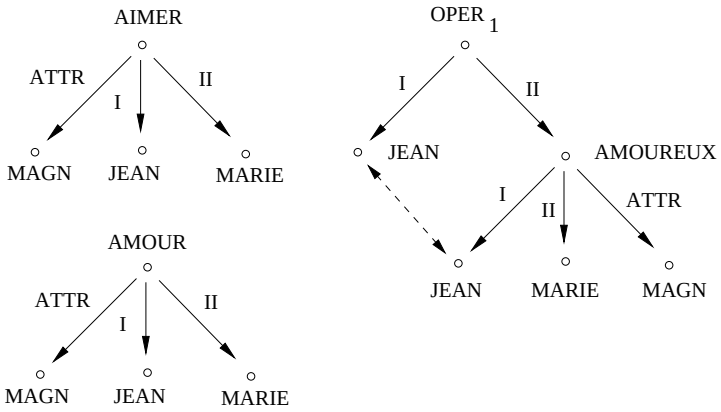

FIG. 2.24 – Structures syntaxiques profondes

AIMER, AMOUR, AMOUREUX, JEAN et MARIE sont les lexies profondes de la représentation. MAGN et OPER_1 sont des fonctions lexicales. Les arcs I et

II correspondent donc, comme nous l'avons dit, aux rôles actanciels. L'arc ATTR exprime une relation attributive.

On voit l'usage qui est fait de la fonction lexicale. Elle incarne syntaxiquement le sens 'très', mais on ne sait pas encore comment ce sens s'exprimera en surface. C'est d'ailleurs un des attraits de cette représentation puisque la subordination de l'item lexical à venir vis à vis de sa base (ici AIMER, AMOUR ou AMOUREUX) est clairement exprimée.

La représentation syntaxique profonde permet de synthétiser une ou plusieurs représentation(s) syntaxique(s) de surface définie(s) comme suit :

La structure est ici encore un arbre de dépendance.

Les nœuds sont des lexies de surface.

Les arcs sont étiquetés par des relations syntaxiques de surface propres à la langue.

Cette structure présente l'ensemble du matériel lexical, y compris les prépositions régies et les collocations, comme le montre la figure (2.25).

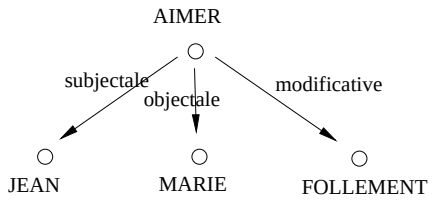

FIG. 2.25 – Structure syntaxique de surface

2.3.2.2 Règles de correspondance

Les transitions d'une représentation à l'autre se font selon des règles de correspondance. Les règles établissant une correspondance entre la représentation semantique et la représentation syntaxique profonde (RSem ⇔ RSyntP) sont les règles sémantiques, celle établissant la correspondance entre la représentation syntaxique profonde et la représentation syntaxique de surface (RSyntP ⇔ RSyntS) sont les règles de syntaxe profonde.

La règle présentée figure (2.26) associe au sémantème 'aimer' les arbres de syntaxe profonde. On voit que les nœuds du graphe et de l'arbre contiennent des variables. Cette règle s'appuie sur le dictionnaire qui stipule que le sens 'X aime Y' peut se lexicaliser (notamment) par AIMER dont les actants syntaxiques profonds réalisent les actants sémantiques. De même, le sens 'X est très' se lexicalise en

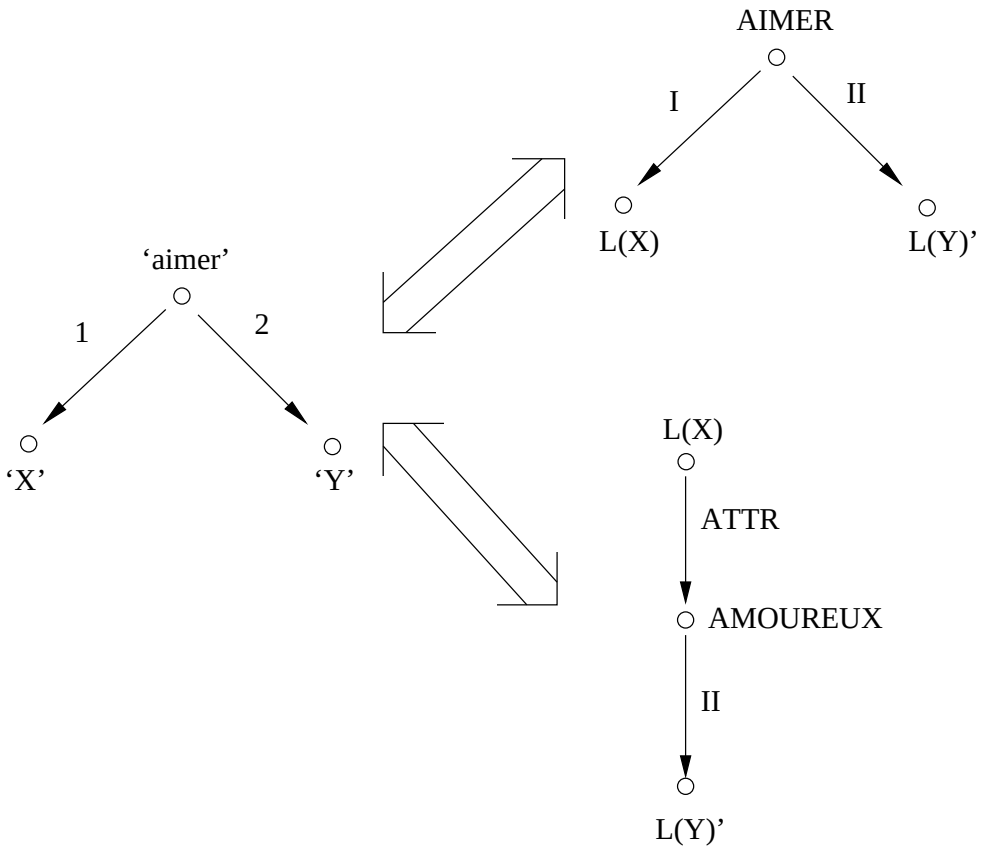

FIG. 2.26 – Rsem ⇔ RSyntP pour 'aimer'

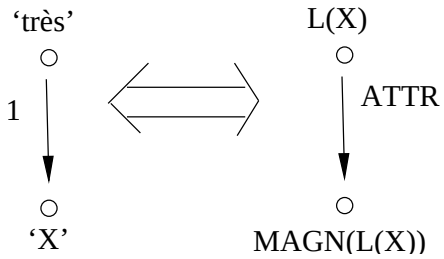

FIG. 2.27 – RSem ⇔ RSyntP pour 'très'

un nœud MAGN(L(X)) relié à l'arbre de syntaxe par une relation attributive (cf. figure(2.27)).

La lexicalisation de MAGN(AIMER) est ensuite réalisée lors de la transition en syntaxe de surface. Ici encore, c'est le dictionnaire qui permet de connaître la valeur renvoyée par la fonction (en l'occurrence *follement*).

On voit qu'une des pierres d'angle de la Théorie Sens-Texte est le dictionnaire. Celui-ci est formalisé et se nomme Dictionnaire Explicatif et Combinatoire. Nous le présentons dans la section suivante.

2.3.2.3 Le Dictionnaire Explicatif et Combinatoire

Le Dictionnaire Explicatif et Combinatoire (DEC) est le dictionnaire concrétisant les principes de la Lexicologie Explicative et Combinatoire (LEC) fondée sur la Théorie Sens-Texte. Il en existe pour l'heure quatre volumes pour le français (cf. (Mel'čuk, 1999)). Comme l'indique son titre, le dictionnaire fournit sur les lexies de la langue une définition sémantique (c'est la dimension explicative) ainsi qu'une recension des collocations et dérivés sémantiques (c'est la dimension combinatoire). Voici un exemple de la structure d'un article, celui de la lexie *risque* $_I$ (article complet dans le DEC IV (Mel'čuk, 1999)) :

La définition est une paraphase synonymique de la lexie définie. Elle spécifie notamment les actants dans une composante principale.

```
Risque de X pour Y =
  fait (fort) dangereux pour Y
  qu'entraîne avec une haute probabilité le fait X
  qui implique Y
```

Le syntactique recence les réalisations syntaxiques profondes des actants sémantiques spécifiés dans la définition.

```
X = I = de N , Aposs, de Vinf
Y = II = pour N
```

Le premier actant peut donc se réaliser de trois façons différentes (*les risques du métier*, *ses risques*, *le risque d'avaler de travers*) et le second n'a qu'une réalisation possible (*les risques pour toi sont importants*).

La zone de combinatoire lexicale énumère les fonctions lexicales associées à la lexie. Chaque fonction lexicale renvoie un ensemble de valeurs. Nous n'avons repris ici que les fonctions les plus caractéristiques.

```
Syn : danger, péril
A1 : risqué1a
Magn : grand, gros | antépos ; considérable, fou, important
AntiMagn : faible < minime < nul
Oper1 : comporter, présenter [ART ~ pour N = Y] ;
        représenter [ART ~]
```

On voit ainsi que le dictionnaire établit des correspondances entre les différents niveaux de représentation. La lexie est liée à un réseau sémantique via sa définition, les équations du type `X = I` indiquent comment les actants sémantiques se réalisent en syntaxe profonde et les régimes et fonctions lexicales donnent les valeurs nécessaires pour la représentation de surface.

Dans la suite du mémoire nous nous réfèrerons essentiellement à un autre dictionnaire, le DiCo (cf. (Polguère, à paraître b)). Il s'agit d'une base de base de données lexicale basée sur la Théorie Sens-Texte. Il se distingue du DEC en ce qu'il ne propose pas de définitions des lexies mais se concentre sur la description des collocations et de la dérivation sémantique. Il comprend en outre un étiquettage sémantique des lexies.

2.3.3 Confrontation des deux modèles

La Théorie Sens-Texte ne relève pas du traitement automatique des langues. Son but n'est pas de construire des textes (d'ailleurs le processus se concentre autour de la synthèse de phrases) mais de proposer un modèle des langues. Cette théorie présente en outre l'intérêt d'attacher une grande importance au lexique avec en particulier une modélisation des collocations. Pour schématiser, on peut dire qu'un Modèle Sens-Texte, tout comme G-TAG, est un module COMMENT-LE-DIRE à la nuance près que la Théorie Sens-Texte ne prévoit pas comment celui-ci doit s'articuler avec un module QUOI-DIRE.

(Polguère, 1998b) propose une perspective pour la lexicalisation en génération en se fondant sur la Théorie Sens-Texte. En opposition avec d'autres approches opérant une lexicalisation précoce (au niveau de la génération profonde) ou bien fondée sur un arbre syntaxique, l'article défend le point de vue d'un module COMMENT-LE-DIRE entièrement dédié aux opérations linguistiques (syntaxiques et lexicales). C'est bien aussi l'objet de G-TAG. L'architecture des deux modules est toutefois assez différente, en raison notamment de la nature de l'entrée.

On a vu que G-TAG prenait en entrée une forme logique. Pour (Polguère, 1998b),

le module QUOI-DIRE fournit un réseau sémantique dont les nœuds sont, comme on l'a vu des sémantèmes. Par conséquent, le QUOI-DIRE n'est pas indépendant de la langue. Cette nuance ne nous semble pas vitale, nous la laisserons donc de côté en considérant pour le moment le réseau sémantique comme une sorte de représentation conceptuelle[5].

Les deux idées phares de la proposition développée dans (Polguère, 1998b) sont les suivantes : d'une part il existe deux phases de lexicalisation (lexicalisation profonde et lexicalisation de surface), d'autre part chaque phase met en œuvre trois types de stratégies. La lexicalisation profonde introduit les unités lexicales profondes et la lexicalisation de surface introduit les unités lexicales de surface. Les unités lexicales profondes sont sémantiquement pleines et requises pour exprimer le sens du message à générer. Les unités lexicales de surface sont, elles, requises pour assurer la bonne formation de la phrase mais ne contribuent pas directement au sens du message. À l'intérieur de chaque phase l'introduction des unités lexicales est soumise à des contraintes de deux sortes définissant trois types d'unités lexicales. (Polguère, 1998b) résume ces options de la façon suivante :

→U : les choix de U sont soumis à des contraintes imposées par le niveau précédent,

→U← : les choix de U sont imposés à la fois par le niveau précédent et pour préparer le suivant,

U← : les choix de U sont imposés uniquement pour préparer le niveau suivant.

La génération profonde introduit six sortes d'unités lexicales dont voici une illustration (LP = Lexies Profondes, FL = Fonctions Lexicales) :

→LP : *et*.

La détermination de cette unité lexicale ne peut être que sémantique dans la mesure où il n'existe pas vraiment de variantes syntaxiques.

→LP← : *aimer*.

La lexie correspond à un contenu sémantique et impose sa catégorie pour la construction syntaxique.

LP← : *fois* (trois **fois**).

La lexie n'exprime pas de sens véritable mais permet de réaliser une certaine structure syntaxique (« il a fait trois éternuements » vs « il a éternué trois fois »).

[5] Bien entendu cette question n'est pas anodine, notamment parce que dans un système de génération il faut bien assurer à un moment ou à un autre l'interface conceptuel / sémantique, cette interface jouant un rôle pour la lexicalisation (cf. (Stede, 1996)).

→FL : **Magn**.

La fonction lexicale exprime un certain sens mais ne contraint pas les choix syntaxiques. Les valeurs renvoyées peuvent modifier un nom, un verbe, un adjectif ou un adverbe.

→FL← : $\mathbf{Real}_{\text{i}}$.

Cette fonction exprime un sens et contraint la syntaxe puisque ses valeurs sont des verbes.

FL← : $\mathbf{Oper}_{\text{i}}$.

Les verbes renvoyés par cette fonction sont vides sémantiquement.

La génération de surface requiert le choix des unités lexicales de surface (Lexies Grammaticales(LG) ou Lexies Collocationnelles(LC)) dont voici des exemples :

→LG : *il*.

L'introduction du pronom n'intervient qu'afin d'exprimer une configuration syntaxique du niveau précédent.

→LG← : *se* (« il **se** regarde »).

Le pronom réfléchi exprime un complément mais est contraint également par le fait que celui-ci ne peut être exprimé littéralement (*« Il regarde lui »).

LG← : *envers* (« l'amour **envers** Jean »).

La préposition est gouvernée par la lexie *amour*.

→LC : *faire* (« **faire** la sieste »).

Le verbe est introduit pour réaliser une valeur de $\mathbf{Oper}_{\text{i}}$.

→LC← : *recevoir* (« **recevoir** une récompense »).

donner *une récompense* ou ***recevoir*** *une récompense* restreignent les choix syntaxiques par rapport à *récompenser* en interdisant d'avoir respectivement le deuxième actant et le premier actant en position de sujet syntaxique.

LC← : *à* (avoir un abcès **à/sur** la jambe vs. **sur** la peau/***à** la peau).

La préposition (ou plus exactement la famille de préposition) est sous-catégorisée par *abcès*, mais la valeur est donnée par la fonction lexicale $\mathbf{Loc}_{\text{in}}(peau)$.

Schématiquement, on peut dire que G-TAG opère aussi une bi-partition des choix lexicaux, sans qu'il y ait pourtant recouvrement exact. Rappelons que la lexicalisation y est opérée en deux temps : on sélectionne à partir du concept une entrée lexicale, puis on sélectionne à partir de cette entrée lexicale l'arbre élémentaire. Les lexies à proprement parler n'apparaissent que dans l'arbre élémentaire, néanmoins le choix d'un verbe ou d'un nom est déjà réalisé par le choix de l'entrée lexicale. Ainsi, les concepts sont associés à des unités lexicales profondes, tandis que les unités lexi-

cales de surface (les prépositions régies, notamment) figurent dans la grammaire TAG.

Une étude plus approfondie montre pourtant des divergences. Dans G-TAG toutes les entrées lexicales peuvent être assimilées à des →LP←[6]. En revanche, G-TAG ne prend jamais appui sur les fonctions lexicales. Est-ce à dire que les textes issus du formalisme ne contiennent aucune collocation ? En fait, le point n'est pas abordé explicitement, mais la base lexicale associée à RWDING dans (Danlos, 1998) contient des entrées lexicales pointant sur des arbres construits autour d'un verbe support : BL(RWDING) = {récompenser, donner-récompense, recevoir-récompense}. Nous verrons plus loin comment l'utilisation des fonctions lexicales $\mathbf{Oper}_i$, $\mathbf{Func}_i$ et $\mathbf{Labor}_{ij}$ permettrait de construire les arbres automatiquement. En revanche, rien n'est prévu pour les →LF← ni les →LF.

Pour ce qui est des unités lexicales de surface, les prépositions régies (LG←) et les verbes support (→LC) figurent dans les arbres élémentaires de la grammaire. Les pronoms (→LG et →LG←) sont générés par le module de post-traitement. Le cas des contraintes syntaxiques liées à l'utilisation d'un verbe support (→LC←) est gérée au niveau supérieur grâce à la base de connaisances indiquant les fonctions syntaxiques des arbres élémentaires de la grammaire. En revanche, le traitement des valeurs de fonctions lexicales choisies pour des raisons strictement syntaxiques (LC←) n'est pas envisagé.

Nous allons donc nous intéresser maintenant à l'introduction des fonctions lexicales $\mathbf{Oper}_i$, **Magn** et $\mathbf{Loc}_{in}$ dans G-TAG.

[6] Comment générer des →LP ou des LP← ? A priori, les →LP correspondent aux connecteurs du discours. Des structures particulières leur sont réservées. Nous n'avons pas examiné le cas des LP← qui semble être assez exceptionnel.

Chapitre 3

Génération des collocations

L'introduction de collocations en génération nécessite de répondre à au moins deux questions :

1 quelles sont les procédures et structures de données permettant d'introduire ces unités lexicales d'un type particulier dans le module COMMENT-LE-DIRE ? Pour notre part, comme nous nous appuyons sur le langage des fonctions lexicales, une partie du travail est déjà en place. Il reste donc à déterminer l'articulation des fonctions lexicales avec le formalisme G-TAG ?

2 quelles sont les informations requises par le COMMENT-LE-DIRE afin qu'il puisse activer une procédure de lexicalisation en tenant compte des éventuelles collocations ? Nous examinerons deux grandes options. La première, que nous appelerons l'« approche conceptuelle », consiste à concevoir un QUOI-DIRE ignorant tout du lexique, en particulier ses caractéristiques collocationnelles. L'apparition de collocatifs n'y est donc pas du tout anticipée. La seconde, que nous appelerons l'« approche lexicale », projette sous forme de concepts l'ensemble des fonctions lexicales syntagmatiques contenues dans le dictionnaire. Un typage est introduit pour indiquer ce qui peut être argument de ces concepts.

3.1 Intégration dans G-TAG

Nous examinons dans cette section la possibilité d'intégrer les fonctions lexicales dans le formalisme G-TAG. Nous nous attacherons à trois fonctions lexicales prototypiques mettant en jeu des interfaces différentes : $\mathbf{Oper}_{\mathrm{i}}$ qui est une fonction purement syntaxique de génération profonde (FL←), $\mathbf{Loc}_{\mathrm{in}}$ qui, lorsqu'elle est sous-catégorisée par une lexie, encode une lexie collocationnelle de surface (LC←)

et **Magn** qui est fonction sémantique de génération profonde (→FL).

3.1.1 FL← : $\mathbf{Oper_i}$

On a vu que G-TAG générait des groupes verbaux à verbes support. Ce que nous proposons ici n'est donc pas vraiment une extension. Toutefois, il est permis d'envisager un nouvel encodage s'appuyant sur la fonction lexicale $\mathbf{Oper_i}$. Nous proposons deux options.

3.1.1.1 1re option

Cette première option consiste à ajouter un nouveau type d'arbres aux bases lexicales. On aura ainsi dans la base lexicale du concept RWDING l'arbre de la figure (3.1). Il comprend l'entrée lexicale (récompense), deux variables pour les arguments à instancier (RWDER et RWDEE) et un nœud d'un type nouveau : $\mathbf{Oper_{12}}$. Ce dernier est une pure information syntaxique. L'entrée lexicale récompense pointe sur une famille d'arbres ancrés par les lexèmes *récompense* et *donner* qui est la valeur de $\mathbf{Oper_{12}}(récompense)$.

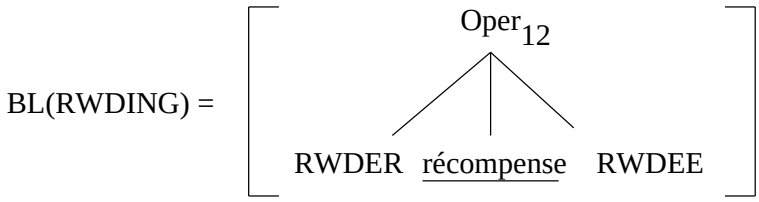

FIG. 3.1 – Arbre de G-dérivation sous-spécifié pour RWDING

L'avantage d'une telle structure est d'expliciter le rôle du verbe vis à vis de *récompense*. Toutefois, elle modifie beaucoup le formalisme. Les arbres de G-dérivation sont en principe des structures proches de la sémantique dont les nœuds sont soit des entrées lexicales sémantiquement pleines soit des variables d'arguments à instancier par d'autres entrées lexicales, également pleines sémantiquement. Or, comme nous l'avons dit $\mathbf{Oper_{12}}$ est un élément purement syntaxique, sa présence dans l'arbre de G-dérivation est donc déplacée.

3.1.1.2 2e option

La seconde possibilité consiste créer une nouvelle famille de type $\mathbf{Oper_i}$ dans la méta-grammaire pour construire les arbres élémentaires. Du coup la base lexicale de

RWDING contient les arbres de G-Dérivation sous-spécifiés de la figure (3.2). Le premier construit nécessairement une phrase (trait `-T,+P`) tandis que le second permet de construire une phrase ou un groupe nominal (trait `-T`).

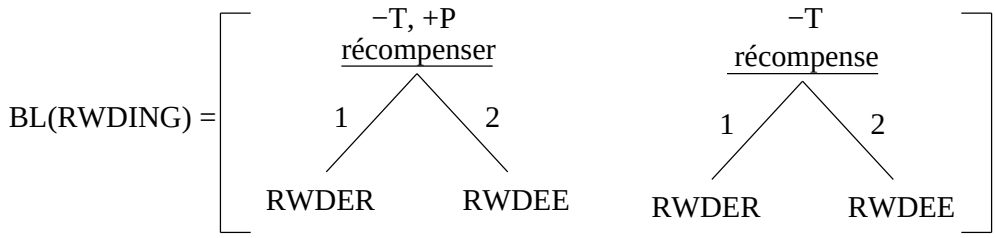

FIG. 3.2 – Arbres de G-dérivation sous-spécifiés pour RWDING

L'entrée lexicale récompense pointe sur la famille d'arbres comprenant notamment les arbres donnés dans la figure (3.3).

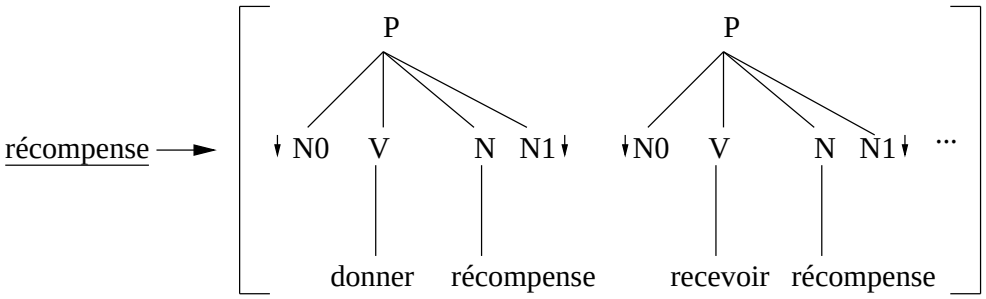

FIG. 3.3 – Arbres élémentaires pour l'entrée lexicale récompense

Cette option change peu le formalisme G-TAG dans la mesure où la fonction lexicale $\mathbf{Oper}_i$ n'y est pas directement intégrée.

3.1.2 LC← : $\mathbf{Loc}_{in}$

Le cas de $\mathbf{Loc}_{in}$ est quelque peu différent. Cette fonction renvoie un ensemble de prépositions régissant le mot-clef et ayant la sémantique de 'se trouvant dans'. Dans le Dictionnaire Explicatif et Combinatoire, cette fonction est fréquemment utilisée pour caractériser le régime d'une lexie nominale. Celle-ci sous-catégorise un groupe prépositionnel dont la préposition est un collocatif du nom ou de la locution nominale sous-catégorisés.

On a par exemple les valeurs suivantes dans le DiCo :

```
ABCES : Y = II = Loc-in N [un abcès sur la joue]
COUP DE BALAI : Y = II = Loc-in N [dans les milieux politiques]
COUP DE SOLEIL : Y = II = Loc-in N [coup de soleil dans le dos
                                    <sur le visage>]
FRISSON : Z = III = Loc-in N [dans le dos, le long des bras, ...]
IRRITATION : X = I = Loc-in N [irritation sur la peau des mains]
MECONTENTEMENT : X = I = Loc-in N | X est un ensemble d'individus
["mécontentement dans la population <chez les infirmiers>"]
```

Une solution possible est d'utiliser des traits avec partage de valeurs. C'est le moyen le plus proche de ce qui est préconisé dans (Polguère, 1998). Les arbres élémentaires correspondants sont présentés sur la figure (3.4). La valeur de la préposition ne sera instanciée qu'après substitution du N.

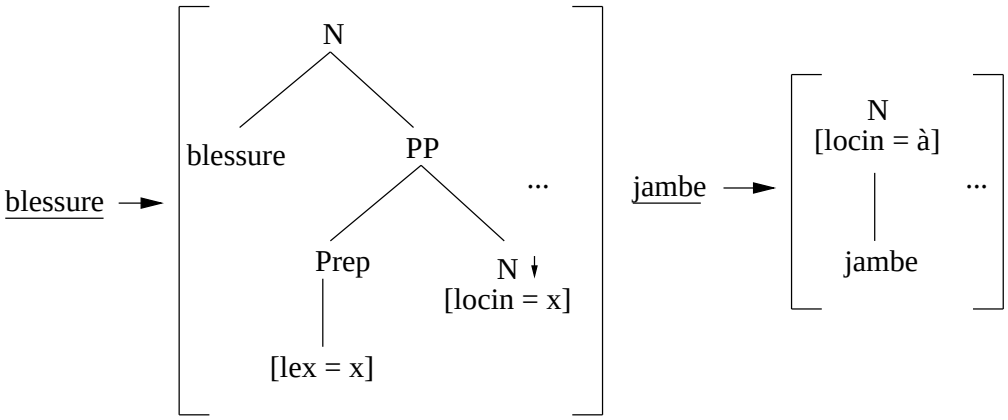

FIG. 3.4 – Arbres élémentaires de blessure et jambe

Si l'on considère que la préposition correspondant à $\mathbf{Loc}_{\text{in}}$ garde une valeur sémantique même lorsqu'elle est régie[1], la représentation proposée fait une entorse au principe TAG qui veut que tout arbre élémentaire ait une seule unité sémantique. Une autre possibilité est donc de disposer des arbres élémentaires donnés figure (3.4). Un nœud supplémentaire (le nœud Locin) est introduit dans l'arbre de G-dérivation.

[1] Pour une étude sur la question de la sémantique des prépositions, cf. (Bonami, 1999)

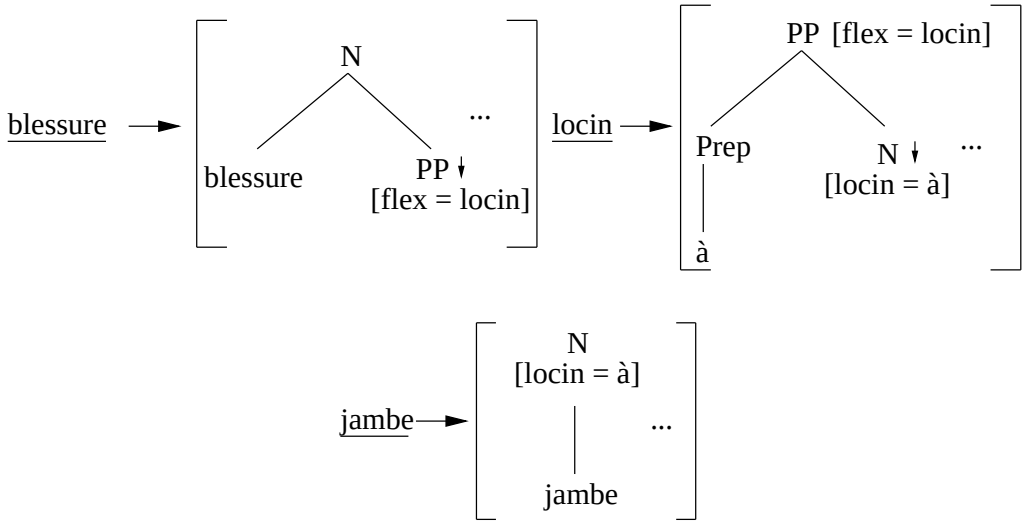

FIG. 3.5 – Arbres élémentaires de blessure, locin et jambe

3.1.3 →FL : Magn

Le formalisme G-TAG n'est pas adapté à la génération des collocations exprimant un sens véritable. La figure (3.6) montre le principe général de lexicalisation dans ce formalisme. On y distingue trois niveaux de représentation :

Niveau conceptuel : chaque concept (C) est interfacé avec plusieurs entrées lexicales (EL).

Niveau lexical : chaque entrée lexicale (EL) pointe vers un ou plusieurs arbres élémentaires (AE).

Niveau syntaxique : les arbres élémentaires (AE) sont ancrés par au moins un lexème. Ils respectent le principe d'unité sémantique.

Pour exprimer une valeur de **Magn**, il nous manque donc un niveau de représentation. En effet, une valeur de **Magn** correspond à une unité sémantique mais dont la réalisation effective est contrainte par la réalisation préliminaire d'une autre unité lexicale. Nous proposons dans les lignes qui suivent deux solutions pour étendre le formalisme.

La façon la plus naturelle de générer les collocations encodées par la fonction lexicale **Magn** est de supposer qu'il existe une relation bi-univoque entre un concept TRÈS et une entrée lexicale magn[2] et d'introduire la fonction lexicale dans l'arbre de

[2] Cette façon de voir est soutenue par l'observation faite en section (1.2.1) que **Magn** encode une valeur très abstraite d'intensification.

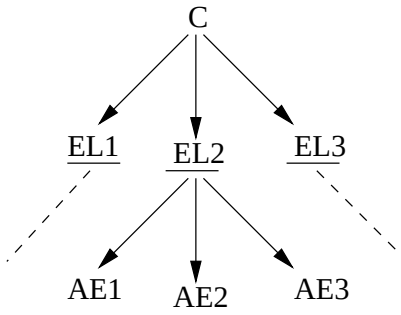

FIG. 3.6 – Lexicalisation en G-TAG

G-dérivation. Nous examinons les questions d'interface conceptuel/sémantique plus loin, section 3.2.2.

3.1.3.1 1re solution

La base lexicale pour le concept TRÈS pointe sur l'entrée lexicale magn comme indiqué figure (3.7). L'entrée magn est associée aux arbres élémentaires de la figure (3.8) qui introduisent un trait de partage de valeur.

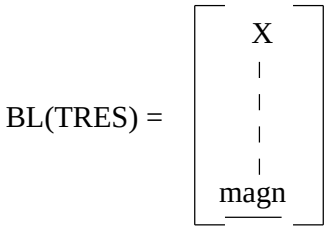

FIG. 3.7 – Base lexicale pour TRÈS

Les arbres élémentaires de l'entrée lexicale magn sont sous-spécifiés, c'est-à-dire qu'ils ne sont pas ancrés lexicalement. L'ancrage effectif est réalisé ensuite par le module de post-traitement. Ceci modifie donc le formalisme TAG puisque ce qui peut venir s'ancrer est soit une unité lexicale, soit un phrasème (comme *à la folie*), donc un arbre.

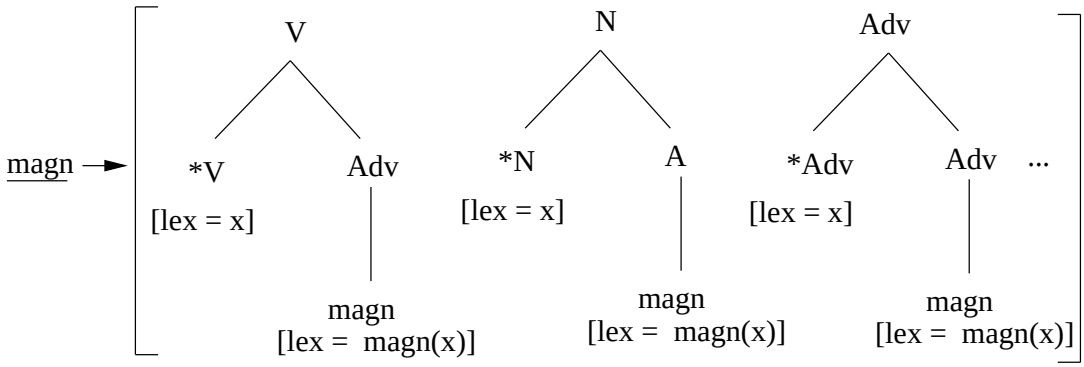

FIG. 3.8 – Arbres élémentaires pour magn

3.1.3.2 2e solution

Une deuxième option consiste à étendre le formalisme en ajoutant des traits aux nœuds des arbres de G-dérivation sous-spécifiés pour les paramétrer. La base lexicale (partielle) de TRÈS est donnée figure (3.9).

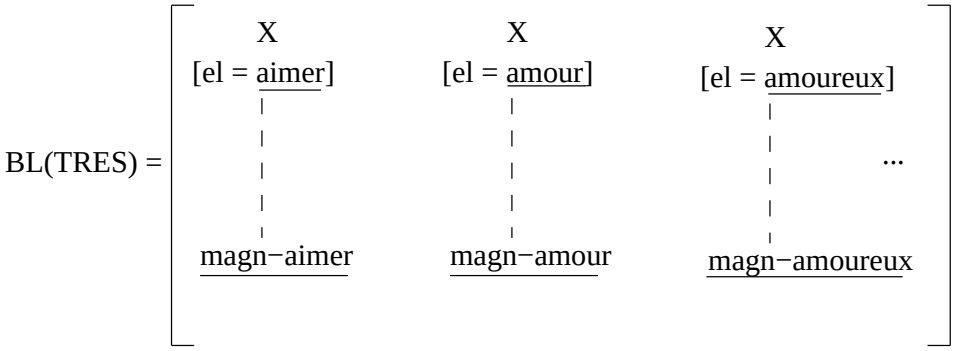

FIG. 3.9 – Base lexicale pour TRES

Chaque entrée lexicale (magn-aimer, magn-amour, ...) pointe sur des familles d'arbres élémentaires TAG classiques.

Remarquons qu'aucune des solutions proposées ne permet de générer *adorer*. Cette unité lexicale correspond à l'expression lexicale de deux concepts, TRÈS et AMOUR. Il faut donc soit envisager une procédure regroupant les concepts TRÈS et AMOUR en un unique concept ADORER, soit être en mesure de lexicaliser une configuration de concepts. Ceci nous conduit donc à revenir à la question de l'entrée du COMMENT-LE-DIRE.

3.2 Interface conceptuel/sémantique

Nous examinons ici deux représentations conceptuelles possibles en entrée du COMMENT-LE-DIRE. L'approche conceptuelle peut être qualifiée de non permissive dans la mesure où l'opération d'intensification doit y être explicite. La seconde approche traite la représentation conceptuelle comme étant une abstraction sur le lexique.

3.2.1 Approche conceptuelle

Cette première approche est soutenue par l'impératif souvent exprimé en génération que le QUOI-DIRE se doit d'être le plus possible indépendant de la langue. Or, comme nous l'avons vu au chapitre 1, les collocations relèvent du savoir lexical. Nous devons donc supposer que le QUOI-DIRE transmette au COMMENT-LE-DIRE une forme logique dont certaines configurations de concepts donnent lieu à une lexicalisation sous forme de collocatif. De ce fait, le QUOI-DIRE ignore tout des collocations, la reconnaissance de celles-ci incombant au COMMENT-LE-DIRE.

On s'interdira par exemple les expressions conceptuelles suivantes :

(1) Jean(x) $\wedge$ $_{e1}$catastrophe(y,x) $\wedge$ très(e1)

(2) Jean(x) $\wedge$ Marie(y) $\wedge$ $_{e1}$éloge(x,y) $\wedge$ très(e1)

Le concept TRÈS doit obligatoirement porter sur un concept graduable. On aura donc plutôt les formules suivantes :

(1') Jean(x) $\wedge$ $_{e1}$qqch(y) $\wedge$ cause(e1,e2) $\wedge$ $_{e2}$csq(z) $\wedge$ $_{e3}$mauvais(z,x) $\wedge$ très(e3)

(2') Jean(x) $\wedge$ Marie(y) $\wedge$ $_{e1}$propos(x,y) $\wedge$ $_{e2}$bon(e1) $\wedge$ très(e2)

Les concepts MAUVAIS et BON étant intensifiables, les formules (1') et (2') sont bien formées[3]. On peut envisager de former les phrases suivantes à partir de ces formes logiques :

(1')a Quelque chose est arrivé à Jean avec de très mauvaises conséquences pour lui.

(1')b C'est une catastrophe épouvantable pour Jean.

(2')a Jean a eu pour Marie des propos très bons.

(2')b Jean a fait l'éloge appuyé de Marie.

Les énoncés numérotés #b comprennent des collocations. Pour les obtenir, il a fallu regrouper des concepts pour les lexicaliser par une unique entrée lexicale. Nous

[3] Il y a toutefois le problème que les prédicats MAUVAIS et BON portent sur des événements réifiés, alors qu'ils devraient porter sur la propriété. Les formules $_{e3}$mauvais(z,x) $\wedge$ très(e3) et $_{e2}$bon(e1) $\wedge$ très(e2) seraient donc mieux exprimées par $_{e3}$mauvais(p,z,x) $\wedge$ très(p) et $_{e2}$bon(p,e1) $\wedge$ très(p) où « p » dénote la propriété du prédicat.

devons donc notamment diposer des bases lexicales suivantes :

1. BL(qqch.+cause+csq+mauvais) → catastrophe
2. BL(propos+bon) → éloge
3. BL(très) → {éloge-magn, catastrophe-magn, ...}

Le concept TRÈS est lexicalisé par une collocation par le biais d'un arbre de dérivation paramétré comme vu section 3.1.3.2.

On peut faire deux remarques à propos de cette démarche. Au plan pratique, tout d'abord, il n'est pas évident de traiter des configurations de concepts. Mais comme on l'a vu, pour générer *adorer* cette ressource est indispensable. D'un point de vue conceptuel, il est peut-être fastidieux de devoir décomposer des objets comme CATASTROPHE ou ÉLOGE. D'une certaine façon, il fait partie de notre savoir conceptuel qu'il existe des événements ayant de mauvaises conséquences pour quelqu'un, ou qu'un propos au sujet de quelqu'un peut être bon ou mauvais. Mais ce savoir n'est pas représenté directement par cette approche.

Nous proposons donc dans la section suivante une autre façon d'envisager la construction d'ontologies.

3.2.2 Approche lexicale

Cette seconde approche établit un lien très fort entre la représentation conceptuelle et le lexique. Nous nous proposons de construire une ontologie à partir du DiCo. Nous justifions dans un premier temps cette démarche puis nous passons à l'examen d'un exemple avec la lexie $abandon_{I.2b}$.

3.2.2.1 Qu'est-ce qu'un concept ?

Projeter une ontologie à partir d'un dictionnaire comme le DiCo revient à considérer d'une part que chaque entrée désigne un concept, d'autre part que les fonctions lexicales (du moins certaines d'entre elles) désignent également des concepts. Ce point de vue est soutenu par le fait que les lexies nous permettent de discrétiser le monde. Les fonctions lexicales correspondent, elles, à ce que l'on peut dire de ces concepts. Les unités lexicales nous servent à parler du monde, et transmettent en même temps une certaine vision du monde. En effet, on rencontre souvent le débat d'une distinction entre le lexique et les connaissances encyclopédiques. On oppose, par exemple, des dictionnaires comme le Robert ou le Larousse sur cette question. On trouvera dans ces deux dictionnaires les mêmes entrées. Mais là où le sens est en

partie exprimé par des illustrations graphiques détaillées, on trouve dans l'autre une énumération de la cooccurence lexicale restreinte. Cette dernière méthodologie a le mérite de fonder le savoir encyclopédique sur le lexique. Toutefois la description n'y est pas explicite puisque les collocations ne sont pas décrites au moyen des fonctions lexicales.

Nous donnons ci-dessous une partie de la cooccurrence lexicale restreinte pour la lexie *bétail* telle qu'encodée dans le DiCo :

$\mathbf{S}_1$: éleveur [de ˜]
$\mathbf{S}_{\text{loc}}$: étable
$\mathbf{Real}_2^{\text{I}}$: élever [ART ˜] ; garder, surveiller [ART ˜] ; faire paître, nourrir [ART ˜] ; faire boire [ART ˜] ; conduire, emmener, mener [ART ˜ $\mathbf{Loc}_{\text{in}}$ N]
$\mathbf{Real}_{\diamond}^{\text{II}}$: abattre [ART ˜]
$\mathbf{S}_{\text{loc}}\mathbf{Real}_{\diamond}^{\text{II}}$: abattoir
$\mathbf{S}_2\mathbf{Fact}_0$: aliment [pour (le) ˜] ; avoine, foin, fourrage, luzerne, maïs, sainfoin, trèfle
$\mathbf{S}_{\text{loc}}\mathbf{Fact}_0$: pacage, pâturage, pâture ; champ, prairie, pré

Le fait que le bétail ait pour fonction d'être abattu (afin d'être consommé) est exprimé lexicalement. On pourra expliquer à un locuteur ignorant le sens d'*abattoir* que c'est le lieu où le bétail se réalise ultimement en tant que bétail.

Pour résumer, nous proposons de projeter les lexies du DiCo au plan conceptuel et de préciser pour chaque concept ce que l'on peut en dire en nous appuyant sur les fonctions lexicales. Cette approche nous semble assez fidèle à la conception du lexique telle qu'elle est envisagée dans la Théorie Sens-Texte. Nous l'illustrons dans la section suivante en projetant un concept ABANDON à partir de la lexie $abandon_{I.2b}$.

3.2.2.2 Exemple : $abandon_{I.2b}$

Nous proposons de projeter un concept ABANDON à partir de l'article du DiCo pour la lexie $abandon_{I.2b}$. Nous tenterons ensuite de générer les phrases suivantes :

1. La maison est à l'abandon
2. Jean a laissé la maison à l'abandon
3. Jean a laissé la maison dans un total abandon

La figure (3.2.2.2) donne l'entrée du DiCo.

```
ABANDON I.2b :
état d'un lieu: ˜ [DU lieu Y] [= QResult(abandon#I.2a)]
```

```
Y = I = ---
{QSyn} délaissement; oubli#2
{Gener} état(1)#2 [d'~]
/*Dans un état d'A.*/
{A1} à [l'~], dans [ART ~ | avec modificateur] //abandonné
/*Complet*/
{Magn} absolu, complet, total
/*Indesirable*/
{AntiBon} triste | antepos
/*Etre dans un etat d'A. [=f1]*/
{Oper1} etre [a l'~]/[dans ART ~ | avec modificateur]
/*Permettre f1*/
{PermOper1} laisser [N=Y a l'~]/
            [dans ART ~ | avec modificateur]
/*Commencer a f1 [=f2]*/
{IncepOper1} tomber [dans ART ~]
/*Permettre f2*/
{PermIncepOper1} laisser aller [N=Y a l'~]/
                 [dans ART ~ | avec modificateur]
Ce bâtiment, laissé à l'abandon depuis 1984,
sera restauré et mis en valeur.
La piece du bas avait le même air de misère et d'abandon.
```

Nous avons effacé certaines informations superflues dans le cadre de cet exposé, mais nous avons conservé les commentaires (placés entre `/*` et `*/`) qui vulgarisent les fonctions lexicales.

En suivant le principe de l'encodage explicite présenté en section 1.2.2.3, nous pouvons déterminer les projections syntaxique et conceptuelle des fonctions lexicales.

$$\textbf{Magn} = \begin{bmatrix} \{\#\}\hat{}\mathsf{Magn} \\ \mathsf{A}[\#\hat{}] \end{bmatrix} \text{ ou } \begin{bmatrix} \{\#\}\hat{}\mathsf{Magn} \\ \mathsf{Adv}[\#\hat{}] \end{bmatrix}$$

Nous tirons de cette représentation trois informations :
- le concept : MAGN(e),
- l'arbre de G-dérivation sous-spécifié : le symbole ˆ indique que celui-ci doit s'adjoindre,
- le type d'arbre élémentaire : adjectival ou adverbial.

Les autres fonctions lexicales projetées ont les encodages explicites donnés en figure (3.10).

$$\mathbf{AntiBon} = \begin{bmatrix} \{\#\}\hat{}\mathsf{AntiBon} \\ \mathsf{A}[\#\hat{}] \end{bmatrix} \text{ ou } \begin{bmatrix} \{\#\}\hat{}\mathsf{AntiBon} \\ \mathsf{Adv}[\#\hat{}] \end{bmatrix} \quad \mathbf{Incep.Oper}_1 = \begin{bmatrix} \mathsf{Incep}[\#] \\ \mathsf{V}[1,\ \#] \end{bmatrix}$$

$$\mathbf{Perm.Oper}_1 = \begin{bmatrix} \mathsf{Perm}[\Omega,\ \#] \\ \mathsf{V}[\Omega,\ 1,\ \#] \end{bmatrix} \quad \mathbf{Perm.Incep.Oper}_1 = \begin{bmatrix} \mathsf{Perm}[\Omega,\mathsf{Incep}[\#]] \\ \mathsf{V}[\Omega,\ 1,\ \#] \end{bmatrix}$$

$$\mathbf{A}_1 = \begin{bmatrix} \{1\}\hat{}\# \\ \mathsf{A}[1\hat{},\ \#] \end{bmatrix} \text{ ou } \begin{bmatrix} \{1\}\hat{}\# \\ \mathsf{Adv}[1\hat{},\ \#] \end{bmatrix} \quad \mathbf{Oper}_1 = \begin{bmatrix} \# \\ \mathsf{V}[1,\ \#] \end{bmatrix}$$

FIG. 3.10 – Fonctions lexicales explicites

On voit que les arguments sémantiques et syntaxiques ne coïncident pas, ce qui pose un problème d'interface sémantique/syntaxe. Nous y reviendrons un peu plus bas. La figure (3.11) montre le concept ABANDON projeté à partir de la fiche du DiCo. La partie haute donne le concept lui-même tandis que la partie basse indique les prédicats qui peuvent le prendre comme argument.

e : ABANDON[x ← LIEU]
MAGN(e)
ANTIBON(e)
PERM(y,e)
INCEP(e)
PERMINCEP(y,e)

FIG. 3.11 – Concept ABANDON projeté à partir de $abandon_{I.2b}$

Un certain nombre de fonctions lexicales ne donnent pas lieu à la création d'un concept, il s'agit des fonctions lexicales syntaxiques.

Les sémantiques associées aux concepts sont[4] les mêmes que celles associées aux fonctions lexicales ayant permis de les projeter :

- ABANDON : 'x est dans l'état-abandon'
- MAGN : 'e est intense'
- ANTIBON : 'e est jugé négativement'
- PERM : 'y cause e'
- INCEP : 'e commence à avoir lieu'
- PERMINCEP : 'y cause que e commence à avoir lieu'[5]

[4] La sémantique de PERM est légèrement simplifiée pour les besoins de l'exposé.

[5] Ce concept devrait être traité compositionnellement, mais cela implique de savoir lexicaliser une configuration de concepts.

En ajoutant les concepts MAISON et JEAN, on peut construire les formes logiques suivantes :

1. maison(x) ∧ abandon(e,x)
2. jean(x) ∧ maison(y) ∧ perm(e,x,e1) ∧ abandon(e1,y)
3. jean(x) ∧ maison(y) ∧ permincep(e,x,e1) ∧ abandon(e1,y)
4. jean(x) ∧ maison(y) ∧ perm(e,x,e1) ∧ abandon(e1,y) ∧ magn(e1)

La figure (3.12) donne les bases lexicales des concepts.

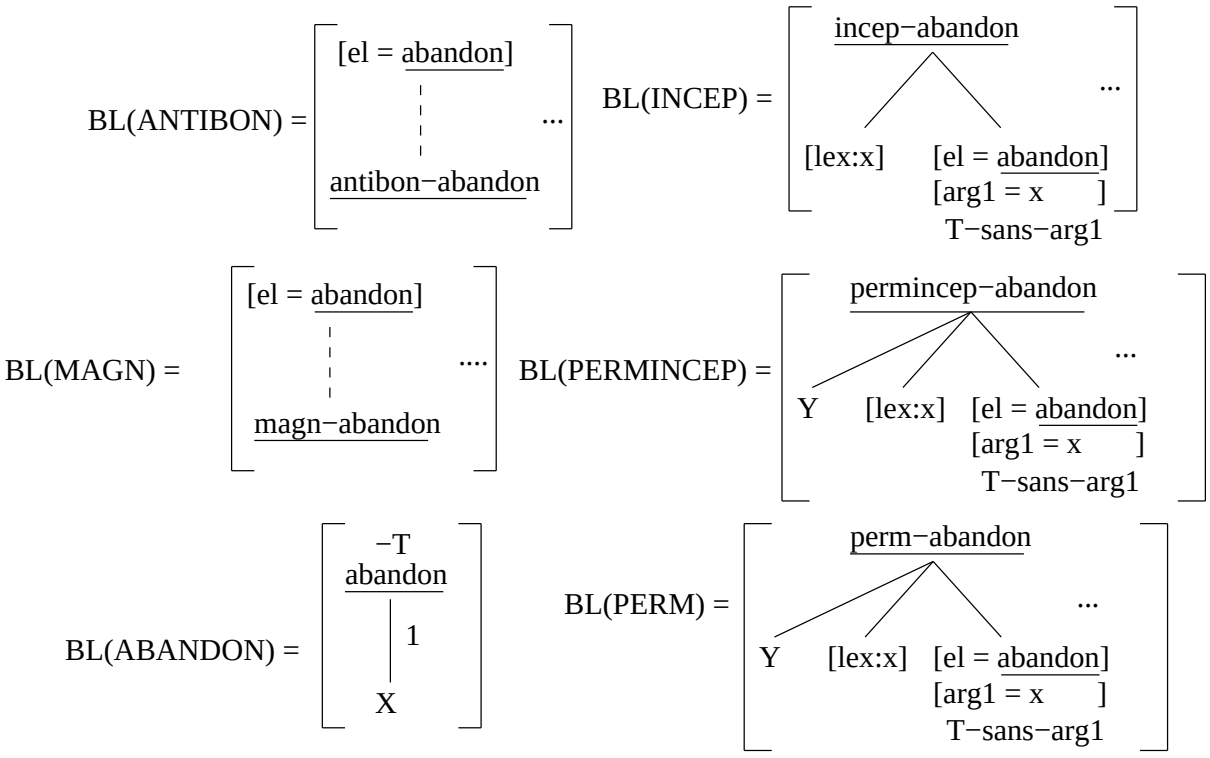

FIG. 3.12 – Bases lexicales

Nous avons ajouté le trait `[T-sans-arg1]` qui sélectionne une construction absolue. Par ailleurs, le partage de valeur permet de récupérer l'argument de abandon afin de le faire remonter. Nous obtenons les phrases suivantes :

(1a) *l'abandon de la maison*,

(1b) *la maison est à l'abandon*,

(2) *Jean a laissé la maison à l'abandon*,

(3) *Jean a laissé aller la maison à l'abandon*,

(4) *Jean a laissé la maison dans un total abandon.*

Nous allons maintenant nous intéresser à la procédure de génération à partir de la forme logique donnée ci-dessus en 4 (cf. figure (3.13)).

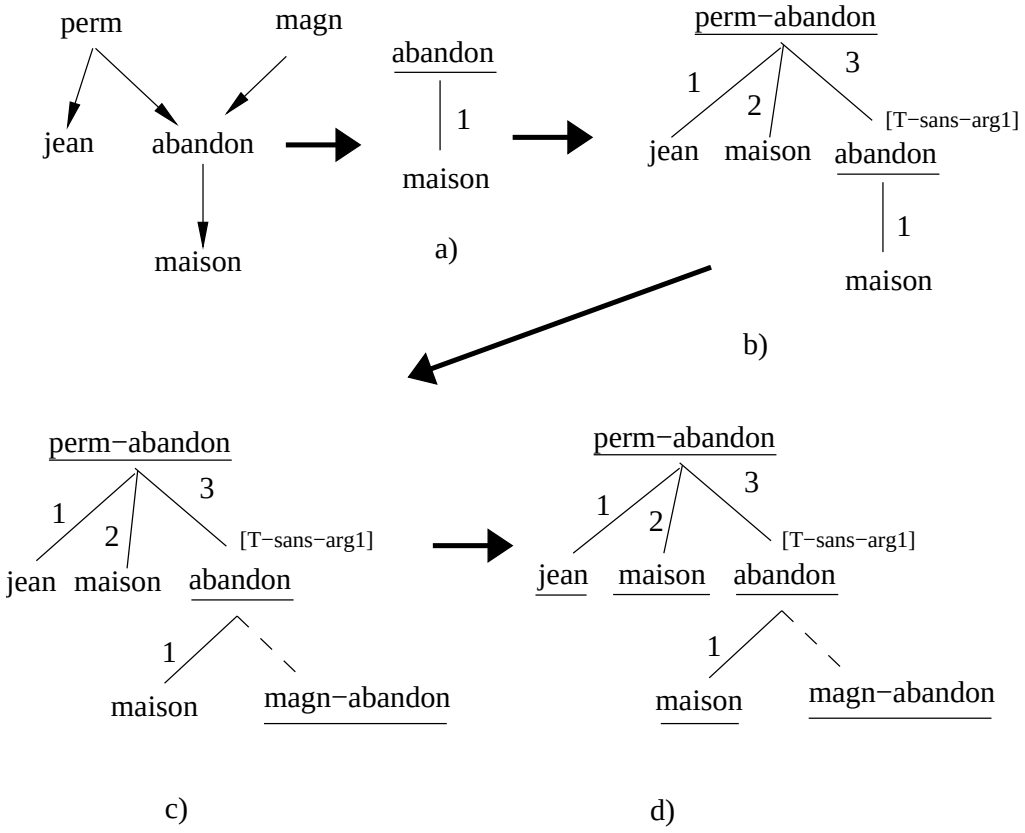

FIG. 3.13 – Construction de l'arbre de G-dérivation à partir de *jean(x)* ∧ *maison(y)* ∧ *perm(e,x,e1)* ∧ *abandon(e1,y)* ∧ *magn(e1)*

L'algorithme pour la construction de l'arbre de G-dérivation est approximativement le même que celui proposé dans (Danlos, 1998) :

> On commence par les prédicats de plus haut niveau. PERM ne peut être lexicalisé car sa base lexicale est paramétrée. On descend les branches et on rencontre alors ABANDON qui est de plus haut niveau que JEAN. Le choix d'un arbre de G-dérivation pour ABANDON est trivial car sa BL ne contient qu'un seul arbre. On obtient la structure donnée en (3.13a) après instanciation de l'argument. La lexicalisation de PERM est désormais possible. Seul l'arbre perm-abandon est disponible. Ses arguments sont instanciés comme indiqué sur la figure (3.13b). De plus l'arbre de G-dériva-

tion sous-spécifié pour perm-abandon ajoute le trait `T-sans-arg1` afin de sélectionner la construction absolue d'*abandon*. MAGN est lexicalisé par l'arbre de G-dérivation paramétré magn-abandon (cf. figure (3.13c)). Enfin les feuilles sont lexicalisées et l'on obtient l'arbre de G-dérivation de la figure (3.13d).

La construction de l'arbre G-dérivé est obtenue en combinant les arbres élémentaires spécifiés dans l'arbre de G-dérivation. La figure (3.14) donne les arbres élémentaires ainsi que l'arbre G-dérivé.

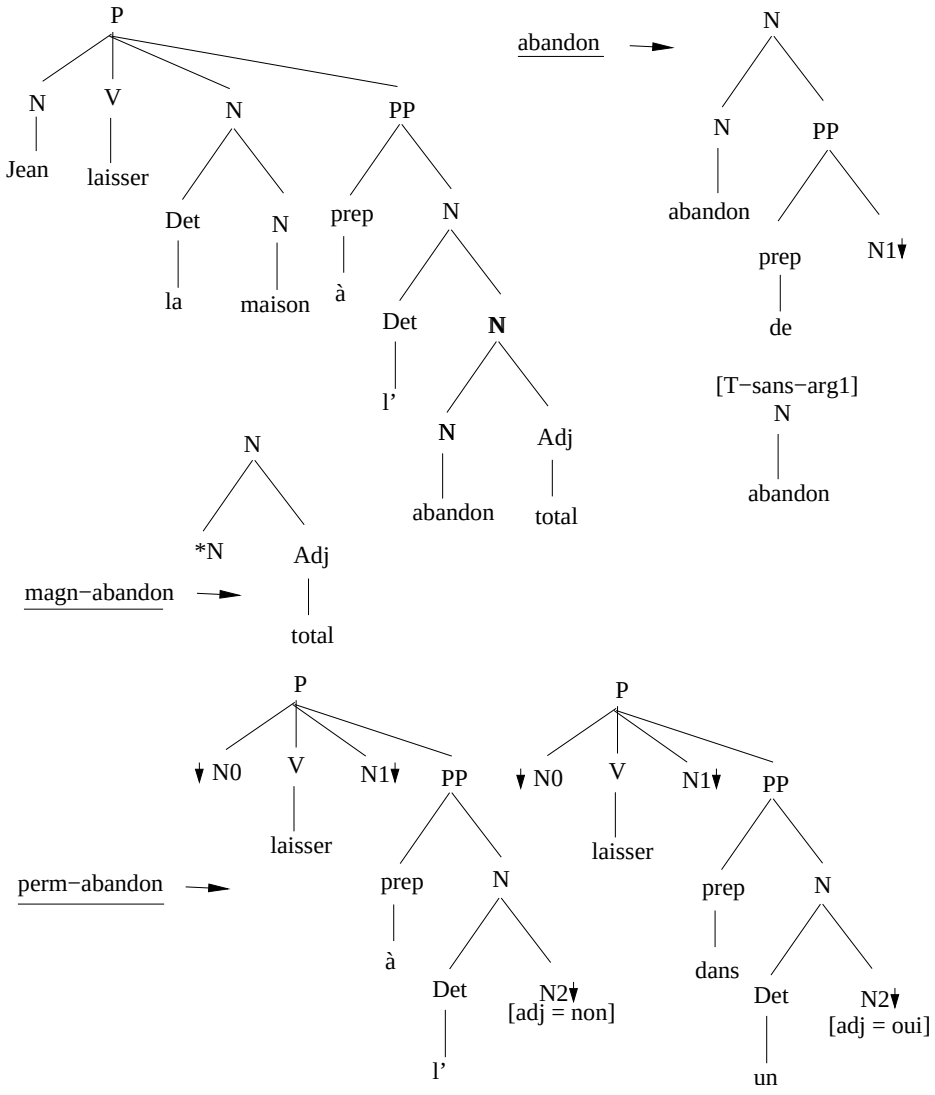

FIG. 3.14 – Arbres élémentaires et arbre G-dérivé final

Conclusion

Le langage des fonctions lexicales s'intègre donc à G-TAG sans qu'il soit nécessaire de modifier profondément ce formalisme. Il reste toutefois des questions.

Le problème de la prise en compte de configurations de concepts (obligatoire, ne serait-ce que pour générer des valeurs fusionnées) n'est pas trivial et doit être examiné en détail.

Nous n'avons pas étudié le problème du choix des valeurs qui réalisent une fonction lexicale. Comme nous avons vu, une fonction lexicale renvoie généralement un ensemble de valeurs. Quelles heuristiques permettraient de les choisir adéquatement ? Pour répondre à cette question, il faut souligner plusieurs points.

Comme nous l'avons signalé en section 1.2.3, les fonctions lexicales ont pour objet de décrire des relations lexicales et non d'autoriser des calculs sémantiques. Ainsi, il n'est pas fait de différence entre « colère noire » et « colère aveugle ». Le collocatif n'est en effet pas toujours seulement l'expression de la sémantique de la fonction lexicale (par exemple 'intense' pour une valeur de **Magn**).

Mentionnons cinq cas de figures de contribution sémantique du collocatif :

- pas de contribution propre : « une colère noire », « une peur bleue »,
- ajout de connotations : « une colère aveugle »,
- différence sur l'échelle de gradation : « des applaudissements forts » vs « des applaudissements frénétiques »,
- contribution sémantique de la valeur : « un abîme infranchissable » vs « un abîme profond »,
- actualisation d'une composante de sens dans la base : « corps à corps sanglant » vs « un corps à corps acharné ».

La notion de gradation est prise en compte dans l'encodage du DiCo. En revanche, rien ne permet de distinguer les valeurs neutres comme *noire* dans « colère noire » de celles qui apportent des éléments de sémantiques comme *aveugle* dans « colère aveugle ». Dans cette dernière famille, il faut également distinguer le simple

apport de connotations (comme c'est le cas avec *aveugle*) d'un apport plus complexe lié au sens de la lexie (comme *infranchissable* dans un « abîme infranchissable »).

Le dernier point souligné relève davantage d'un problème d'encodage, comme remarqué dans (Kahane, à paraître). Certains **Magn** peuvent effectivement être paraphrasés de façon plus pertinente par des fonctions lexicales complexes ou par des configurations incluant les fonctions standards $\mathbf{Real}_{\mathrm{i}}$ ou $\mathbf{Fact}_{\mathrm{i}}$ mettant ainsi en évidence l'actualisation d'une composante. De façon générale, la mise en jeu de composantes propres ou standard reste à étudier.

Bibliographie

Abeillé A., *Les nouvelles syntaxes*, 1993, Armand Colin.

Aït-Kaci H. et Nasr R., "*Login : A Logic-Programming Language with Built-In Inheritance*", in *Journal of Logic Programming*,3, 1986.

Beauchesne J., *Dictionnaire des cooccurrences*, 2001, Guérin.

Bonami O., *Les construction du verbe : le cas des groupes prépositionnels argumentaux*, Thèse, 1999, Université de Paris 7

Danlos L., *Génération automatique de textes en langues naturelles*, 1985, Masson.

Danlos L., "*GTAG, un formalisme lexicalisé pour la génération de textes inspiré de TAG*", 1998, T.A.L. 39.

Danlos L. et Roussarie L., "*Génération automatique de textes*", in Pierrel J.-M.,*Ingéniérie des langues*, Hermès, 2000.

Danlos L., Gaiffe B. et Roussarie L. "*Document Structuring à la SDRT*", in Proceedings of the 8th European Workshop on Natural Language Generation, Toulouse, pp. 11-20, 2001.

Kahane S., "*Grammaires de dépendance formelles et théorie Sens-Texte*", 2001, Actes TALN'2001a.

Kahane S. et Polguère A., "*Un langage formel d'encodage des fonctions lexicales et son application à la modélisation des collocations*", 2001b, in B. Daille et G. Williams, *Collocations*, Actes Journée Atala.

Kahane S. et Polguère A., "*Formal foundations of lexical functions*", 2001, Workshop on Collocation, ACL 2001c.

Kahane S., "*Une blessure profonde dans le Dictionnaire Explicatif et Combinatoire. Sur le lien entre les définitions lexicographiques et les fonctions lexicales*", à paraître, .

Mel'čuk I., *Cours de morphologie générale*, vol. 1, 1993, Presses universitaires de Montréal.

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

Mel'čuk I., *Vers une linguistique Sens-Texte. Leçon inaugurale*, 1997, Paris : Collège de France.

Meunier F., *Implantation du formalisme de génération G-TAG*, Thèse, 1997, Université de Paris 7, Talana.

Mel'čuk I. et al., *Dictionnaire Explicatif et Combinatoire*, volume IV, 1999, Presses Universitaires de Montréal.

Polguère A., "*La Théorie Sens-Texte*", 1998a, Dialangue, Vol. 8-9, Université du Québec à Chicoutimi.

Polguère A., "*Pour un modèle stratifié de la lexicalisation en génération de texte*", 1998b, T.A.L. 39.

Polguère A., "*Lexical Functions Standardness*", à paraître a), in *Festschrift in Honour of Igor Mel'čuk*, John Benjamins.

Polguère A., "*Étiquettage sémantique des lexies dans la base de données DiCo*", à paraître a), T.A.L. 44 n°2.

Reiter E. & Dale R., *Building Natural Language Generation Systems*, 1999, Cambridge University Press.

Roussarie L., *Un modèle théorique d'inférence de structures sémantiques et discursives dans le cadre de la génération automatique de textes*, Thèse, 2000, Université de Paris 7, Talana.

Shieber S. et Schabes S., "*An alternative Conception of Tree Adjoining Derivation*", *Computationnal Linguistics*, vol. 20 n°1, 1994.

Stede M., "*Lexical options in multilingual generation from a knowledge base*", in G. Adorni, M. Zock (eds.) : *'Trends in Natural Language Generation - An Artificial Intelligence Perspective'* (Papers from the 4th European WS on Natural Language Generation) Springer, Berlin/Heidelberg, 1996.