CHAPITRE II

LES CORPUS ARBORES

Nous montrons dans une première section les notations employées pour rendre compte des relations syntaxiques et nous rappelons la nature des phénomènes à noter. Nous présentons dans une deuxième section un corpus arboré, ***Susanne***, qui représente une réalisation exemplaire par la finesse de l'annotation produite et par la manière dont les choix effectués sont documentés. La troisième section est consacrée à l'utilisation de corpus arborés et de parseurs pour l'étude de la phraséologie. La dernière section examine les enjeux théoriques de corpus arborés et les conditions pratiques de leur emploi.

1. DIVERSITE DES CORPUS ARBORES

Même si les jeux d'étiquettes varient et si les séquences à catégoriser, selon qu'on regroupe ou non des unités polylexicales, il est relativement aisé de se faire une idée d'un corpus étiqueté : une catégorie est associée à chaque occurrence du texte. Cette belle simplicité disparaît dès qu'on aborde les corpus arborés, c'est-à-dire « décorés » d'arbres. L'annotation résultante peut varier du tout au tout. Il s'agit en effet de délimiter des groupes, de les nommer (les catégoriser), et de statuer sur leurs relations. À ces trois niveaux, les points de vue sont multiples. Tous les constituants ne font pas l'unanimité : c'est le cas du syntagme verbal, cher à la tradition chomskyenne, et rejeté par M. Gross. G. Sampson (Sampson, 1995, p. 4) cite à ce propos une expérience significative. À la rencontre de 1991 de l'Association for Computional Linguistics, des chercheurs en TALN appartenant à neuf institutions différentes se sont vu demander de délimiter les constituants d'un ensemble de phrases. Pour l'exemple suivant, voici les seuls parenthésages qui ont fait l'unanimité :

He said this constituted a [very serious] misuse [of the [Criminal Court] processes].

Nous définissons les principales facettes des corpus arborés : les notations disponibles, la manière d'obtenir les analyses, les types d'analyses et d'analyseurs, les niveaux d'annotation syntaxique.

1.1 Noter des relations syntaxiques

1.1.1 Arbres, graphes et relations

Les arbres sont le dispositif habituel pour noter les relations syntaxiques. La tradition veut que les feuilles soient à la base et la racine au sommet. On distingue les *nœuds terminaux*, les feuilles et les autres nœuds, appelés *non-terminaux*. Ces nœuds non-terminaux englobent les *nœuds pré-terminaux*, qui dominent directement les feuilles. Si l'on considère un nœud et ses fils, un arbre matérialise deux relations particulières : celle de dépendance immédiate, entre le père et ses fils, et celle de précédence immédiate, entre un nœud aîné et son ou ses cadet(s).

A un nœud est associée une étiquette (SN) et éventuellement des « décorations » : une série d'associations trait-valeur du type {genre=masculin, nombre=singulier}. Comme pour l'étiquetage morpho-syntaxique, les étiquettes simples ou complexes se ramènent en fait toutes à une structure de traits : SN = {cat=SN} et SNMS = {cat=SN, genre=masculin, nombre=singulier}. Conceptuellement, à chaque nœud, correspond donc non une étiquette mais une structure de traits.

Les deux relations de dépendance et de précédence ne suffisent pas à noter la variété des phénomènes syntaxiques. Deux nœuds frères séparés par d'autres nœuds peuvent constituer une unité discontinue (comme la négation complète *ne ... pas* dans *Ne me quitte pas*). L'anaphore suppose un lien entre l'anaphorique et son antécédent : il s'agit d'un lien entre des nœuds qui ne sont généralement pas frères, mais à des niveaux différents de la représentation syntaxique. Certains constituants sont « flottants » : leur insertion à un endroit donné ne rend pas compte de leur portée réelle. C'est le cas des adverbes de phrase comme heureusement dans *Heureusement Jean a terminé son année* et *Jean a heureusement terminé son année*. L'attachement réel d'un nœud peut rester en suspens, même pour un locuteur : c'est le cas dans *Jean a heureusement terminé son année*, où *heureusement* peut modifier la phrase dans son ensemble, mais également le syntagme verbal seul (*il est heureux que Jean ait terminé son année* / *Jean a terminé son année d'une manière heureuse*).

Visiblement l'arbre ne suffit plus à noter tous ces phénomènes. On peut souhaiter recourir à des graphes moins limités, où un nœud peut être le point d'arrivée de plusieurs arêtes. Il faudrait même que ces graphes puissent être « polychromes »[1] pour visualiser aisément les diverses

[1] Cf. (Marandin et Cori, 1993) pour une proposition formelle en ce sens.

relations à l'œuvre. Une autre direction de travail consiste à utiliser des descriptions logiques d'arbres, où l'on ne manipule ni des arbres ni des graphes, mais la conjonction logique des divers types de relations identifiées entre les nœuds. Elle est explorée par Vijay-Shanker (1992), dans la lignée des travaux de M. Marcus (Marcus *et al.*, 1983).

Cette remise en cause de l'arbre comme mode fondamental de notation syntaxique n'est pas nouvelle[2]. Elle peut plus profondément renvoyer au choix entre grammaires de constituants et grammaires de dépendances.

1.1.2 Grammaires de constituants et grammaires de dépendance

On trouve dans Tesnière les prolégomènes des grammaires de dépendance. I. Mel'cuk, qui s'inscrit dans cette lignée, contraste (1988, p. 12-42) les grammaires de dépendance avec les grammaires de constituants (*phrase structure grammars*). Les grammaires de constituants mettent au premier plan l'inclusion d'un segment dans une catégorie syntagmatique et des segments d'un type dans des segments de niveau supérieur (deux constituants sont ou bien enchâssés ou bien disjoints). La plupart des nœuds y sont non-terminaux. Les nœuds d'un niveau donné sont ordonnés linéairement. Les relations de domination sont entre constituants et non pas entre mots. Les grammaires de dépendance révèlent les liens hiérarchiques entre mots. Tous les nœuds sont terminaux. Ils ne suivent pas forcément un ordre linéaire. Un arbre de dépendance du type [V sont reçus [N [N Pierre][Coord et][N Jacques]] ne contient aucune information directe concernant l'ordre linéaire des mots dans l'énoncé, qui peut se réaliser sous la forme *Pierre et Jacques sont reçus* comme sous la forme *Sont reçus Pierre et Jacques*.

Ce sont les grammaires de constituants qui sont majoritairement employées pour les corpus annotés syntaxiquement. La langue traitée peut expliquer le choix fait. Les grammaires de constituants semblent mieux adaptées aux langues à ordre des mots relativement contraint et aux syntagmes nettement identifiables, comme l'anglais[3]. Les grammaires de dépendance conviennent davantage aux langues où l'ordre des mots est plus libre (le finnois, par exemple). Contribuent sans doute également à cette prépondérance le poids des travaux proprement linguistiques qui relèvent de cette tradition mais aussi le fait que la technologie des parseurs pour les langages informatiques fait aussi appel aux grammaires hors contexte. Les grammaires de dépendance offrent cependant l'avantage de faciliter l'utilisation des relations hiérarchiques entre mots d'un énoncé. Si l'on veut dégager les cadres de sous-catégorisation des verbes, par exemple, cette approche permet un élagage immédiat qui ne

[2] On trouve dans le modèle GPSG (Gazdar *et al.*, 1985) la volonté de découpler dans les règles hors contexte la relation de dominance et l'ordre linéaire, c'est-à-dire la précédence.

[3] Toutefois, le parseur ENCG - English Constraint Grammar (Karlsson *et al.*, 1995), crée des structures de dépendance pour l'anglais. (Karlsson, 1994, p. 130-142) fournit plusieurs exemples de résultats commentés (extraits d'un manuel informatique, d'*Alice au pays des merveilles* et d'une encyclopédie). Inversement, certains formalismes cherchent à rendre compte des variations d'ordre des mots dans le cadre des grammaires de constituants.

conserve que les liens de dépendance pertinents.

1.1.3 Notations textuelles

Puisque les arbres constituent la notation prépondérante, nous continuons à parler de corpus arborés. Le stockage d'arbres pour leur traitement informatique suppose de passer d'une représentation dans le plan à une représentation textuelle essentiellement linéaire : elle figure par l'enchâssement la relation de dépendance et par la succession la relation de précédence. Des dispositifs annexes permettent de dépasser les limites des arbres. Il s'agit généralement d'indices attachés aux nœuds et de renvois à ces indices pour exprimer les autres relations.

Le format de présentation des corpus arborés varie. Il peut être horizontal : c'est le cas de cet exemple[4] emprunté à la banque d'arbres d'IBM France :

> [N Ce_DEDEMMS guide_NCOMS N][V [P leur_PPCA6MP P] permet V_VINIP3 [P de_PREPD [Vi se_PPRE6MP familiariser_VPRN [P avec_PREP [N les_DARDFP opérations_NCOFP [P de_PREPD [N réseau_NCOMS [A local_AJQMS A] N] P][A effectuées_VTRPSFP [P par_PREP [N les_DARDMP utilisateurs_NCOMP N] P] A] N] P] Vi] P] V] ._.

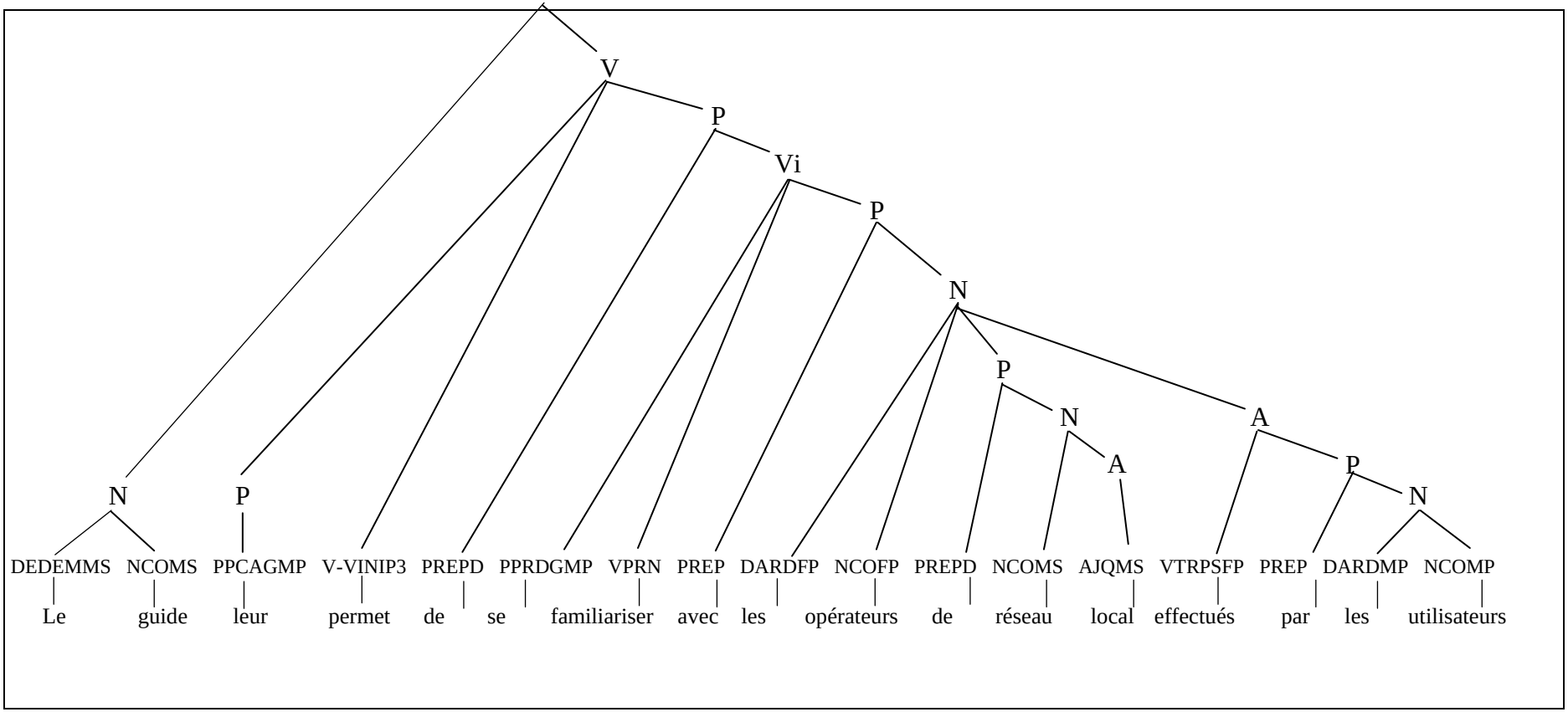
L'étiquette du constituant est souvent fournie deux fois : au début et à la fin du groupe en cause, probablement pour faciliter le repérage visuel des groupes et des frontières. Les enchâssements font apparaître une hiérarchie, dont l'indentation, plaçant les constituants de même niveau à une même distance de la marge gauche, facilite la perception :

```
[N Ce_DEDEMMS guide_NCOMS N]
[V
  [P leur_PPCA6MP P]
     permet V_VINIP3
     [P de_PREPD [Vi se_PPRE6MP familiariser_VPRN [...]
```

Il peut également être vertical. On distingue comme dans ***Susanne*** formes, étiquettes de mots, parties d'arbres. Pour l'exemple choisi :

[4] Cité dans (Leech *et al.*, 1996, p. 6).

Ce	DEDEMMS	[N .
guide	NCOMS	. N]
leur	PPCA6MP	[V [P . P]
permet.	V_VINIP3	.
de	PREPD	[P .
se	PPRE6MP	[Vi .
familiariser	VPRN	.
avec	PREP	[P .
les	DARDFP	[N .
opérations	NCOFP	.
de	PREPD	[P .
réseau	NCOMS	[N .
local	AJQMS	[A . A] N] P]
effectuées	VTRPSFP	[A .
par	PREP	[P .
les	DARDMP	[N .
utilisateurs	NCOMP	. N] P] A] N] P] Vi] P] V]
._.		


V
P
Vi
P
N
P
N
A
P
N
A
N
P
DEDEMMS NCOMS PPCAGMP V-VINIP3 PREPD PPRDGMP VPRN PREP DARDFP NCOFP PREPD NCOMS AJQMS VTRPSFP PREP DARDMP NCOMP
Le guide leur permet de se familiariser avec les opérateurs de réseau local effectués par les utilisateurs

Le mot figure en première colonne, sa catégorie en seconde. La troisième colonne fournit une partie de l'arbre syntaxique : le point y marque l'insertion du sous-groupe constitué de la catégorie et du mot. Les deux premières lignes correspondent ainsi au sous-arbre [N [DDEMMS Ce][NCOMS guide] N].

Ces deux présentations, verticale et horizontale, correspondent à l'arbre donné dans la figure ci-contre[5] (nous le simplifions en omettant les catégories pré-terminales).

1.2 Obtenir des analyses

Il est possible d'associer à un texte des annotations syntaxiques plus ou moins complexes de manière purement manuelle. Mais, sauf à disposer de moyens humains et matériels très importants, cela limite la taille du texte ainsi analysé. C'est le choix qui a été fait pour ***Susanne*** (cf. section 2), parce qu'il s'agissait d'obtenir une analyse aussi fouillée que possible. C'est encore le cas des corpus qui sont balisés à la main pour servir de corpus d'apprentissage de grammaires probabilistes (cf. chapitre VIII), comme celui développé en commun par l'université de Lancaster et IBM (Eyes et Leech, 1993). Dans ce cas (*ibid.* p. 132), à l'opposé de ***Susanne***, il s'agit d'insérer des arbres dits « squelettiques » (parenthésage et catégorisation des constituants principaux). L'autre possibilité est l'analyse syntaxique automatique, ou parsage, mieux adaptée au traitement de gros volumes textuels.

Entre « travail manuel » et parsage, bien des intermédiaires existent : l'intervention humaine peut se produire en amont (pour délimiter des groupes ou éliminer des catégories parasites) ou en aval pour trancher entre plusieurs analyses : c'est le cas du système TOSCA (Halteren et Oostdijk, 1993, p. 154) ou pour améliorer l'analyse produite : c'est la solution retenue par ***Penn Treebank*** (Marcus *et al.,* 1993).

1.3 Types d'analyse

1.3.1 Analyse partielle / analyse complète

L'analyse peut être partielle ou complète. Complète : c'est un arbre qui couvre l'ensemble de la phrase, dont les feuilles sont les mots de la phrase. Partielle : à une phrase donnée correspond(ent) un ou plusieurs arbre(s) qui laisse(nt) des parties qui ne sont pas analysées.

Une analyse partielle peut correspondre à l'incapacité du parseur, pour une phrase particulière ou en général, à produire des structures qui couvrent l'intégralité des données analysées. Mais une analyse partielle

[5] *A priori*, il est toujours possible de passer automatiquement d'un format à un autre, et d'en fournir une version réellement arborée comme ici, même si le détail du codage propre à tel corpus peut rendre difficile la mise au point du traitement nécessaire.

peut correspondre aussi au fait de ne s'intéresser qu'aux composants d'une certaine nature syntaxique. C'est ainsi qu'en terminologie automatisée, les extracteurs de groupes nominaux se concentrent sur ces syntagmes, où figurent les dénominations polylexicales du domaine. Dans la phrase suivante de ***Mitterrand1***[6] : « le Louvre , libéré du le[7] ministère des les finances , cela représente un immense palais , le plus grand musée du le monde — un kilomètre sept cent si vous voulez en faire le tour — imaginez la fatigue des les pieds des les visiteurs : il faut que les œuvres d' art soient quand même à la portée de ceux qui veulent se déplacer », sont retenus par LEXTER (cf. 3.4), à partir de la version lemmatisée, les groupes nominaux suivants :

[SN [SAdj [Adj immense]][SN [Nom palais]]]

[SN [SAdj [Adj grand]][SN [SN [Nom musée]][SP [Prep de][SN [Det [Art le]][SN [Nom monde]]]]]]

[SN [SN [Nom fatigue]][SP [Prep de][SN [Det [Art le]][SN [SN [Nom pied]][SP [Prep de][SN [Det [Art le]][SN [Nom visiteur]]]]]]]

[SN [SN [Nom œuvre]][SP [Prep de][SN [/Nom art]]]]

Une analyse partielle peut enfin avoir pour but de produire une version simplifiée de la phrase, en laissant de côté des composants ou des parties de composants conçus comme secondaires. Par exemple, le parseur peut extraire l'association sujet – verbe – complément d'objet, et ignorer les compléments circonstanciels, si l'objectif est d'étudier la sous-catégorisation des verbes, leurs cadres syntaxiques et leurs arguments typiques.

1.3.2 Une seule analyse ou plusieurs

Le résultat peut fournir, pour un segment donné, une seule analyse ou plusieurs.

On distingue deux types d'ambiguïtés. Ambiguïtés réelles : un locuteur ne pourrait pas trancher. Hors contexte, par exemple, il est difficile de savoir comment analyser *état de l'art abstrait* (*Cette thèse commence par un [état de l'art] abstrait* / *Ce critique d'art présente l'état de l'[art abstrait]*). Ambiguïtés techniques : le savoir dont dispose le parseur n'est pas suffisant pour choisir entre des possibles[8], mais un locuteur n'a pas de difficultés à le faire, en fonction de ses connaissances générales ou au vu du contexte[9]. C'est le cas des rattachements prépositionnels et

[6] Emission de TF1 *Ça nous intéresse, Monsieur le Président*, du 28 avril 1985.
[7] Dans le pré-traitement, les contractions préposition + article défini (*aux*, *du*, *des*) sont décomposées pour faciliter les opérations ultérieures.
[8] T. Briscoe (1994, p. 99) donne l'exemple de la définition de *youth hostel (A hostel for usu. young people walking around country areas on holiday for which they pay small amounts of money to the youth hostels association or to the international yha)* dans le *Longman Dictionary of Contemporary English* (*LDOCE*). Le parseur inclus dans Alvey Natural Language Tools, avec un dictionnaire de 20 000 entrées, a produit plus de 2 500 analyses. Voir (Souter et Atwell, 1994, p. 151) pour un autre exemple d'analyse ambiguë.
[9] À l'inverse, un annotateur confronté à des phrases isolées peut se trouver dans l'incapacité de trancher (Black *et al.*, 1993, p. 40).

adjectivaux. Dans l'expression *traitement du langage naturel*, s'il ne dispose pas dans son lexique de l'expression *langage naturel*, un analyseur peut ne pas savoir s'il faut rattacher *naturel* à *traitement* ou à *langage*.

Voici, à titre d'exemple, les pourcentages d'ambiguïté obtenus par le système TOSCA sur un corpus d'1,5 million de mots de prose anglaise contemporaine (Halteren et Oostdijk, 1993, p. 155) :

Nombre d'analyses différentes	fiction	non-fiction
1	22 %	20 %
2	15 %	15 %
3-5	17 %	19 %
6-10	15 %	15 %
11-20	10 %	12 %
21-100	15 %	16 %
> 100	6 %	3 %

Ces chiffres donnent une idée des difficultés rencontrées en analyse syntaxique automatique.

1.3.3 Sous-spécification

Il est possible de laisser une analyse sous-spécifiée, c'est-à-dire incomplète sur un point donné. Cela revient à limiter artificiellement l'ambiguïté, en la laissant implicite. Par exemple, les attachements prépositionnels ou adjectivaux, souvent difficiles à effectuer automatiquement, peuvent être " laissés en suspens " pour permettre une post-édition spécifique. Le parseur ENGCG (Voutilainen et Heikkila, 1994, p. 190) dans *fat butcher's wife*, indique juste que *fat* s'attache à un nom à droite sans décider s'il s'agit de *butcher* (*la femme du gros boucher*) ou de *wife* (*la grosse femme du boucher*) et n'effectue pas non plus les rattachements des adverbiaux, notoirement délicats. C'est encore le cas du parseur Fidditch (Hindle, 1994) dans ***Penn Treebank*** qui ne rattache pas les groupes dont il ne peut pas déterminer avec certitude le rôle dans une structure de plus haut niveau (cf. chapitre VIII). Cela peut aboutir à fournir pour une phrase une suite d'arbres non reliés entre eux. Dans certains cas, des nœuds sont laissés sans étiquette quand leur délimitation est claire, mais pas leur catégorie (Black *et al.,* 1993, p. 19).

1.4 Analyseurs de texte « tout-venant »

Nous précisons les types de parseurs qui sont effectivement employés pour l'annotation de vastes corpus, ainsi que les choix qui conditionnent leur fonctionnement : production d'une seule analyse ou de plusieurs,

analyse descendante ou montante.

Certains formalismes syntaxiques contemporains comme LFG, HPSG, les grammaires d'arbres adjoints (Abeillé, 1993) ou comme le modèle Gouvernement et Liage ont donné lieu à la réalisation de parseurs. Toutefois, ces analyseurs sont avant tout destinés à tester le traitement par ces formalismes de phénomènes linguistiques complexes (dépendances à distance, etc.). S'ils visent à avoir la « couverture » la plus large possible, il faut entendre cet objectif comme la capacité à traiter un à un la plupart des problèmes syntaxiques d'une langue et non comme la capacité à traiter l'enchevêtrement de ces problèmes dans des phrases authentiques longues et complexes, qui peuvent même violer certaines « règles » grammaticales. Les parseurs de ces obédiences ne semblent pas dans l'immédiat utilisables sur de vastes corpus[10]. À notre connaissance, il n'existe d'ailleurs pas de corpus annoté selon leurs principes. Par opposition aux parseurs avant tout destinés à tester des formalismes syntaxiques raffinés, l'objectif des analyseurs qui sont évoqués dans ce chapitre est le parsage robuste. Il s'agit, pour reprendre les critères[11] de F. Karlsson (1994, p. 122), de pouvoir analyser, sans se bloquer, du texte « tout-venant », (en fournissant éventuellement des résultats partiels), d'aboutir à un taux satisfaisant d'analyses correctes[12] (*i.e.* où les mots sont dominés par une étiquette syntaxique unique et adéquate) et de ne pas aboutir à des résultats aberrants pour des phrases de longueur et de complexité « raisonnable ». D. Hindle (1994, p. 105) rejoint cette caractérisation. Il insiste en outre sur le fait que le parseur doit toujours produire « quelque chose », même sur un énoncé non grammatical. Il tient, mais c'est un point qui ne fait pas l'unanimité, à ce qu'un résultat et un seul soit retourné pour une phrase donnée. Il souhaite enfin que le parseur permette une amélioration incrémentale.

Les langages artificiels (langages de programmation, langages de représentation de connaissances) sont conçus *a priori* pour éviter toute ambiguïté : quand un programme est exécuté, son comportement, à un moment donné de son exécution, avec des données déterminées, doit être univoque. L'ambiguïté est au contraire centrale pour les langues naturelles. Elle est souvent ressentie comme une difficulté pour les traitements automatiques. Beaucoup de parseurs pour les langues naturelles ont pour visée la production de l'ensemble des analyses possibles. Ce peut être le cas au niveau de la phrase dans son ensemble, comme dans le système TOSCA. Ce peut être aussi le cas en analyse partielle. Certains analyseurs, en revanche, visent à ne fournir qu'une seule analyse. C'est le cas de Fidditch (Hindle, 1994), utilisé pour ***Penn Treebank***. Cette deuxième possibilité, à l'évidence, facilite la production de gros volumes de texte arboré, puisque le post-traitement manuel n'a pas à trier parmi les possibles.

L'objectif d'une ou de plusieurs analyses complètes pour du texte tout-

[10] Certains chercheurs pensent même que ces modèles avant tout théoriques sont de peu de profit pour développer des analyseurs utilisables, au contraire des grandes grammaires descriptives (Black *et al.*, 1993, p. 77).

[11] Nous ne reprenons pas son exigence de rapidité, pour des raisons expliquées au chapitre VIII.

[12] F. Karlsson (*ibid.*) cite l'objectif, qui paraît extrêmement ambitieux de 90 % d'analyses justes. Cf. les pourcentages d'ambiguïté fournis en 1.3.2.

venant est encore loin d'être réalisable. Les parseurs capables de produire des résultats partiels sont donc nécessaires, ce qui favorise les analyseurs montants. Les analyseurs montants (*bottom-up*) regroupent progressivement des structures de niveau de plus en plus élevé, les analyseurs descendants (*top-down*) suivent une approche inverse : des niveaux supérieurs vers les mots. Les premiers sont plus appropriés que les seconds pour fournir des résultats partiels : en quelque sorte, ils « savent » s'arrêter en chemin, en produisant des groupes qui ne sont pas forcément tous reliés, mais qui peuvent déjà être utilisés.

1.5 Niveaux d'analyse

L'examen des corpus arborés existants permet dans (Leech *et al.*, 1996, p. 9) de distinguer, par ordre de complexité croissante, les niveaux d'annotation suivants[13], illustrés sur l'exemple utilisé *supra* :

1.5.1.1 Simple parenthésage des constituants

Ce sont en fait des crochets qui sont le plus souvent utilisés :

[Ce guide][[leur] permet [de [se familiariser [...]

1.5.1.2 Étiquetage des constituants

C'est la représentation fournie plus haut (dans cet exemple, seules les étiquettes des nœuds pré-terminaux sont plus complexes).

On appelle *parsage squelettique* (*skeleton parsing*) le fait de s'en tenir à ces deux niveaux, voire au premier seul. Ce « dégrossissage syntaxique », qui peut être effectué manuellement à relativement faible coût, peut suffire à certaines analyses automatiques ultérieures (recherche de cadres de sous-catégorisation) ou servir de base d'entraînement à un analyseur probabiliste (cf. chapitre VIII).

1.5.1.3 Indication des relations de dépendance

Elle fournit les liens entre les « gouverneurs » (Tesnière ou Mel'cuk) ou « têtes » et leurs « dépendants »[14]. Leur notation se fait par des flèches. Ces liens relient uniquement des mots, à la différence des grammaires de constituants, où les ensembles reliés peuvent correspondre aussi bien à des mots qu'à des groupes de mots.

Nous empruntons les notations du parseur ENGCG (Voutilainen et Heikkila, 1994) pour illustrer cette approche sur notre exemple (> indique que la tête est à droite, la première des deux catégories suivant l'arrobas,

[13] D'autres informations sont distinguées pour un corpus d'oral transcrit et les caractéristiques syntaxiques propres à l'oral : réitérations, faux démarrages, etc. Nous ne les présentons pas, puisque nous avons fait le choix de ne traiter que les corpus d'écrit.

[14] Nous suivons ici la terminologie de (Mel'cuk, 1988, p. 23). La dénomination *dépendant* y est préférée à celle de *modifieur*, parce qu'elle est plus générique.

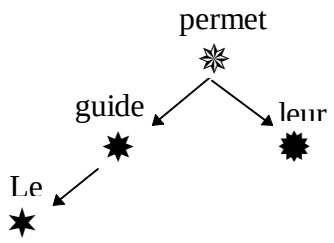
@, renvoie au mot examiné, la seconde au mot tête) :

```
Ce          @DN>
guide       @NV2>
leur        @PV>
permet
[...]
```

@DN> signifie que *Ce* est un Déterminant dépendant du premier Nom à droite (si c'était le deuxième, la notation serait @DN2>). Une autre notation, indiquée dans (Leech *et al.*, 1996, p. 26) assortit chaque mot d'un numéro d'ordre à sa gauche et éventuellement à droite du numéro de la tête dont il dépend :

```
1    Ce       D   2
2    guide    N   4
3    leur     P   4
4    permet   V
[...]
```

ce qui correspond à :



Le mot 1 (*Ce*) dépend du mot 2, qui, comme le mot 3, dépend du mot 4. Ce dernier, qui est la « tête », ne dépend de rien. Il est encore possible (*ibid.*, p. 27) de représenter un graphe de dépendance par une expression parenthésée où chaque parenthèse ouvrante est suivie d'une tête, puis des dépendants de celle-ci, et ce de manière récursive[15] :

```
[V permet [N guide [D Ce]][P leur][...]
```

1.5.1.4 Indication des relations fonctionnelles

Il s'agit de noter les fonctions comme sujet, objet direct, objet indirect etc. :

```
[N <Sujet> Ce_DEDEMMS guide_NCOMS N][V [P <ObjetIndirect> leur_PPCA6MP P]
permet V_VINIP3 [...] ._.
```

1.5.1.5 Classification plus fine des syntagmes

Elle peut être assurée par un système de traits : [N{genre=masc, nombre=sing} Ce_DEDEMMS guide_NCOMS N][V{mode=indicatif, temps=présent, personne=3} [P{nombre=plur} leur_PPCA6MP P] permet V_VINIP3 [...] ._.

1.5.1.6 Relations " logiques " ou profondes

Il s'agit d'indiquer les liens de co-référence, de rassembler les constituants discontinus. Dans le cas présent, un indice (entre chevrons) peut manifester la coréférence entre leur et le sujet implicite (explicité par

[15] Du moins dans les cas où il n'y a pas de discontinuités.

un constituant vide) de se familiariser :

> [N Ce_DEDEMMS guide_NCOMS N][V [P <8> leur_PPCA6MP P] permet V_VINIP3 [P de_PREPD [N <8> N] [Vi se_PPRE6MP familiariser_VPRN [...] ._.

Ces constituants vides peuvent servir ensuite à faciliter le repérage des relations prédicat / arguments dans les phrases (Marcus *et al.,* 1993, p. 321).

1.5.1.7 Information sur le rang d'une unité syntaxique

Le niveau d'enchâssement des constituants est ajouté (il peut le plus souvent être calculé en fonction du niveau de parenthésage).

2. UNE REALISATION EXEMPLAIRE : *SUSANNE*

Susanne est un sous-ensemble de ***Brown*** qui avait déjà été manuellement analysé à Gothenburg. Il comprend 64 extraits de 2 000 mots chacun, soit 128 000 mots, relevant de quatre des genres distingués par ***Brown*** : reportage journalistique, Belles Lettres, écrit scientifique et technique, aventure et fiction. Le corpus obéit à un format vertical, comme nous l'avons vu au chapitre précédent, avec un mot par ligne, et dans l'ordre la référence, le statut (correction ou non), la catégorie pré-terminale, le mot, son lemme, et l'analyse syntaxique.

2.1 Une annotation « exhaustive »

Nous choisissons de présenter en détail ce corpus arboré « manuellement » pour trois raisons.

En premier lieu, c'est l'un des plus faciles d'accès, gratuitement et sans formalités.

En second lieu, le schéma d'annotation est l'un des plus documentés qui soit (Sampson, 1995) : les choix faits sont discutés en détail, ils sont exposés dans des documents aisément accessibles. Cela permet de comprendre et d'utiliser pleinement le résultat : « Les conventions d'annotation de ***Susanne*** proposent une méthode pour représenter tous les aspects de la grammaire anglaise qui sont suffisamment définis pour être susceptibles d'une annotation formelle. Les catégories et les limites entre elles sont spécifiées de manière suffisamment détaillée pour que, dans l'idéal, deux analystes annotant indépendamment le même texte et se référant aux mêmes conventions soient forcés de produire la même analyse structurale » (Sampson, 1994, p. 169).

Enfin, ***Susanne***, comme le souligne l'acronyme : Surface and Underlying Structural ANalyses of Natural English, vise une annotation

aussi exhaustive que possible (pratiquement tous les niveaux définis *supra* y sont représentés) (*ibid.* p. 170) : « son but (comparable à celui de la taxonomie de Linné au dix-huitième siècle dans le domaine de la botanique) n'est pas d'identifier les catégories qui sont optimales sur le plan théorique ou qui reflètent nécessairement l'organisation psychologique de la compétence linguistique des locuteurs, mais simplement d'offrir un schéma de catégories et des façons de les utiliser qui rende aisé aux chercheurs en TALN l'enregistrement systématique et sans ambiguïté de l'usage réel, sans malentendus sur des emplois locaux d'une terminologie analytique*. »* En ce sens, ***Susanne***, qui résulte d'une annotation entièrement humaine, explore les limites de l'annotation syntaxique. Nombre des annotations que ce corpus fournit ne pourraient pas être ajoutées automatiquement à d'autres corpus, au moins dans l'immédiat. En disposer, de façon expérimentale et sur un corpus de taille réduite, permet cependant d'évaluer l'intérêt de chacune d'entre elles pour les recherches, tant linguistiques que computationnelles.

2.2 Informations fournies dans Susanne

Voici les choix faits pour ***Susanne*** aux différents niveaux d'analyse définis *supra*.

352 étiquettes sont utilisées pour l'étiquetage des mots. Sampson (1995) fournit pour les catégories fermées la liste exhaustive et pour les catégories ouvertes les critères d'attribution. Les noms propres sont répartis en noms de personne, noms de lieux etc.

Les nœuds portent jusqu'à trois types d'information : catégorie, fonction et indice (permettant de relier le nœud à un autre nœud).

Les relations fonctionnelles suivantes sont indiquées : sujet logique, objet direct logique, objet indirect logique, agent du passif, sujet de surface, objet de surface, circonstants de lieu, de direction, de temps, de manière etc.

Les étiquettes catégorielles fournissent de nombreuses informations sur les constituants ainsi nommés (forme et type de verbe pour les groupes verbaux par exemple).

Des indices lient les paires de nœuds pour montrer l'identité référentielle entre des constituants qui se trouvent dans certaines configurations syntaxiques. Une étiquette spécifique dans le champ réservé au mot représente la « trace *»* : c'est-à-dire la position logique d'un constituant placé en fait ailleurs ou qui est effacé dans la structure syntaxique de surface. Simultanément, un constituant « déplacé *»* porte une autre étiquette marquant ce déplacement et un indice le lie à la « trace *»* correspondant à sa position « logique *».* Dans l'exemple suivant[16], *John wanted to go* :

[Nns:s123 John] wanted [Ti:o s123 to go]

[16] (Leech *et al.*, 1995, p. 19)

:s indique la fonction sujet, :o la fonction objet de l'infinitive (Ti) pour le verbe *wanted*. Le « fantôme » s123 indique la position logique du sujet de surface *John*. L'indice 123 établit le lien entre la réalisation de surface et le « fantôme ».

Les conventions de notation des étiquettes des nœuds permettent de distinguer les étiquettes pré-terminales, celles des syntagmes, celles des propositions et celles des unités « racines ».

3. PHRASEOLOGIE ET TRAITEMENTS SYNTAXIQUES

Les corpus arborés sont disponibles depuis le début des années quatre-vingt dix, c'est-à-dire depuis moins longtemps que les corpus étiquetés, accessibles depuis les années quatre-vingts. La primauté de l'anglais se fait ici écrasante : il n'existe pas à ce jour de corpus arboré du français aisément disponible[17].

En TALN, ces corpus servent surtout à la mise au point des parseurs. L'observation de corpus arborés permet de préciser les règles à employer, d'analyser automatiquement des corpus de taille plus importante, de retravailler les règles en jeu et ainsi de suite. Cette utilisation est évoquée au chapitre VIII. Les corpus arborés servent également d'étapes vers des traitements sémantiques (cooccurrences syntaxiques et similarités). Le chapitre IV traite cet aspect.

Les recherches linguistiques qui ont recours à des corpus arborés sont donc encore rares. Nous centrons notre analyse sur le traitement de la dimension phraséologique du langage, pour la langue générale — ce sont les « expressions figées », les « mots composés » mais surtout en langage de spécialité — ce sont les termes. C'est une zone à la lisière de la syntaxe et du lexique (Corbin, 1992). Nous présentons des utilisations de corpus arborés et d'analyseurs « robustes » pour rendre compte, en français et en anglais, de ces fonctionnements langagiers.

3.1 Le renouveau des études linguistiques de la phraséologie

Les expressions toutes faites, comme les noms composés (*un champignon atomique*), les verbes composés (dans des constructions à verbe support comme *mettre en évidence*), les locutions adverbiales (*à la volée*), prépositionnelles (*à la fin de*) ou conjonctives (*à seule fin que*), ont souvent été reléguées aux marges des traitements lexicographiques[18]. D'abord, ces unités polylexicales s'insèrent malaisément dans les

[17] Le Centre Scientifique d'IBM France a cependant développé au début des années quatre-vingt dix un corpus arboré de 400 000 mots (débats en français du parlement canadien, manuels IBM) qui peut être acheté. Nous en donnons un exemple *infra*.

[18] En français du moins. Les dictionnaires d'expressions idiomatiques foisonnent pour l'anglais.

dictionnaires sur support papier[19]. Où faire figurer *champignon atomique*, sous l'entrée *champignon* ou sous *atomique* ? Le rattachement à *champignon* paraît naturel, toutefois, c'est bien d'énergie nucléaire qu'il s'agit, et on souhaiterait maintenir ce lien. Où faire entrer *à la volée* ? Ces locutions sont d'ailleurs soumises à déformation (la réalisation originelle *goulet d'étranglement* est concurrencée par *goulot d'étranglement*), mais si les dictionnaires déconseillent certaines variantes, ils ne répertorient pas pour autant toutes les variantes effectives. Ensuite, on voit souvent dans ces séquences la partie « imagée *»,* métaphorique de la langue, comme le souligne A. Rey (Rey et Chantreau, 1979, p. I-XIII), ce qui conduit alors à privilégier une étude de l'origine et de l'évolution de ces séquences et peut-être à sous-estimer leur place dans la langue courante : « un dictionnaire de locutions, s'il n'est pas un simple recueil de traductions, ne peut être qu'historique *»* (*ibid.,* p. XII). Enfin, les limites de l'ensemble considéré sont floues, et variables les critères qui permettent de dire qu'une séquence fonctionne comme un « mot composé ». Si l'on considère *verre à vin* comme un nom composé, faut-il en faire de même de toutes les séquences similaires : *verre à cognac*, *verre à apéritif*, *verre à kyr* ... ?

La maîtrise de ces « mots en plusieurs mots *»* est pourtant essentielle dans l'apprentissage d'une langue. Ils s'avèrent en effet souvent opaques dans la phase de compréhension et causes d'hésitations dans la phase de production. C'est pourquoi Mel'cuk leur donne une place centrale dans son *Dictionnaire Explicatif et Combinatoire du Français*. Ses fonctions lexicales (Mel'cuk, 1988) visent à mettre au jour les réalisations lexicales les plus probables des mots pour exprimer une modification sémantique donnée. Le degré fort se dit ainsi *à chaudes larmes* quand il s'agit de *pleurer* et *à tout rompre* quand le verbe est *applaudir*.

Depuis une quinzaine d'années, la phraséologie suscite un renouveau d'intérêt en linguistique ainsi qu'en TALN. Dans la lignée logique des études menées sur les possibilités combinatoires des mots simples, qui soulignaient les multiples restrictions existantes (Guillet, 1990), les études du LADL ont montré l'importance des « mots composés *».* Elles ont abouti en particulier à un dictionnaire électronique des mots composés en français (Silberztein, 1993). Ce dictionnaire constitue un inventaire extrêmement poussé des expressions, sur le plan quantitatif, mais aussi sur le plan qualitatif. Chaque entrée est assortie de la description de ses variantes possibles. En TALN, l'évolution des formalismes vers la lexicalisation, c'est-à-dire la réduction des règles « générales *»* au profit de règles rendant compte des particularités d'emploi des mots sinon un par un, du moins par classes réduites, s'est accompagnée d'un renouveau des études et des propositions de traitement des expressions dites figées[20].

L'étude des unités polylexicales a conduit un certain nombre d'auteurs (Gazdar *et al.,* 1985 ; Abeillé, 1993 ; Habert et Jacquemin 1995) à postuler que ces unités relèvent des règles générales de la grammaire,

[19] Il n'en va bien sûr pas de même pour un dictionnaire électronique. Les fonctions de recherche permettent de séparer l'entrée concernée et les points d'accès.

[20] Cf. (Abeillé, 1993) pour une présentation sur ce point dans trois formalismes contemporains.

mais qu'elles obéissent à des contraintes supplémentaires[21], et qu'en particulier elles sont moins flexibles que les syntagmes libres de même catégorie : par exemple, on ne peut dire en conservant le même sens *#champignon très atomique*[22] ou *#champignon atomique et dangereux*, etc. Dans la logique de cette approche, on peut examiner une séquence qui constitue éventuellement une unité polylexicale, étudier les transformations syntaxiques dont elle est passible, et en tirer un constat global sur le « degré de figement » de cette séquence. L'hypothèse est que, plus une séquence est figée, c'est-à-dire moins elle accepte de transformations syntaxiques, plus il y a de chances qu'il s'agisse d'une unité polylexicale. C'est l'hypothèse défendue par G. Gross (1988).

L'apport des corpus à ce double renouveau porte sur deux points. En premier lieu, étant donné une expression jugée « contrainte » quant à ses possibilités de transformation, les corpus permettent de chercher si ses réalisations effectives confirment ce jugement. C'est ce que nous examinons en 3.2 et en 3.3 pour des expressions de la langue générale et des termes techniques, respectivement. Deuxièmement, l'ensemble des unités polylexicales est par définition ouvert. C'est par ce biais notamment que s'enrichit le lexique, en particulier dans les domaines techniques et scientifiques. L'observation des corpus sert alors à accroître le lexique des expressions. C'est ce que nous montrons pour les langages de spécialité en 3.4.

3.2 La flexibilité en corpus d'expressions polylexicales

H. Barkema (1993, 1994) se fixe pour objectif la « mesure » de la flexibilité réelle, en corpus, d'expressions toutes faites. Il examine donc les variations, c'est-à-dire les suites de mots qui sont apparentées à ces expressions et qui résultent d'une transformation graphique, phonétique, morphologique ou syntaxique (*gagner le cocotier* pour *gagner le coquetier* résulte d'une approximation phonétique, par exemple). Certaines de ces variations constituent des variantes, c'est-à-dire des équivalents effectifs de l'expression en cause (*infarctus myocardique* pour *infarctus du myocarde*, par exemple).

3.2.1 Les variations en corpus d'expressions « toutes faites »

Pour effectuer le repérage de telles variations, Barkema (1994) recherche les occurrences d'expressions courantes et les suites de mots qui en sont proches dans un vaste corpus, celui de Birmingham, qui rassemble 20 millions de mots. Ce corpus fournit par exemple 111 occurrences

[21] (Barkema, 1993) s'inscrit dans la même vision de hiérarchies de contraintes, tout comme, dans un autre cadre (van der Linden, 1992).

[22] Comme dans (Gazdar *et al.*, 1985) et (Barkema, 1994, p. 42, note 8), le # signale que la séquence en cause est grammaticale mais qu'elle ne peut pas être interprétée « idiomatiquement ». Elle pourrait dénoter un champignon fortement irradié et ne peut pas renvoyer au nuage caractéristique d'une explosion atomique.

inchangées de l'expression *cold war*[23] (*guerre froide*) ainsi que les 13 exemples suivants qui en constituent des variations :

1	renewed Cold War
2	the melting Cold War
3	the world Cold War
4	continuing, ever-present 'cold' war
5	the Cold War won by Europeans who 'destalinized' Eastern Europe
6	the cold war which threatened to divide the world into two ideological armed camps
7	a not-so-cold war against Kaddafi
8	the awkward cold war thought up by the American paranoids, who should be back in the law offices of middlewestern towns
9	a period of cold and hot civil war which ended with Hitler's invasion of Austria
10	a kind of cold civil war
11	the cold war that existed between the two giants, the United States and ...
12	the Cold War in Washington
13	the cold war between the Nature Conservancy Council and the farmers

Barkema répartit variations et emplois non modifiés selon le schéma syntaxique auquel ils obéissent :

Schéma	occurrences et numéros
[{déterminant} cold war]	111 occ.
[{déterminant} {adjectif} cold war]	3 occ. (1, 2, 4)
[{déterminant} cold war {proposition}]	2 occ. (6, 11)
[{déterminant} cold war {syntagme prépositionnel}	2 occ. (12, 13)
[{déterminant} cold war {participe passé}]	1 occ. (5)
[{déterminant} {adjectif} cold war {participe passé}]	1 occ. (8)
[{déterminant} Adv cold war {syntagme prépositionnel}]	1 occ. (7)
[{déterminant} {nom} cold war}	1 occ. (3)
[{déterminant} cold {adjectif} war}	1 occ. (10)
[{déterminant} cold {coordonnant} {adjectif} {adjectif} war {proposition}]	1 occ. (9)

3.2.2 " Mesurer " la flexibilité

Après cette première étape de recueil, Barkema se fixe pour objectif d'évaluer, et même de « mesurer » la flexibilité observée. Les variations effectives de la séquence dans un corpus jugé représentatif sont-elles

[23] L'étude précise de cette séquence s'inscrit dans une recherche plus vaste : l'examen des variations de 450 expressions dans le même corpus (Barkema, 1993).

prévisibles ? Au contraire, sont-elles plus importantes ou moins importantes que ce à quoi on pouvait s'attendre ?

L'hypothèse sous-jacente est que la flexibilité dépend au premier chef du schéma syntaxique de départ de la séquence examinée. Pour pouvoir porter un jugement sur ces variantes observées, c'est-à-dire déterminer si *cold war* est aussi flexible qu'on pourrait s'y attendre, il faut d'abord caractériser la flexibilité effective du schéma sous jacent : [{adjectif} {nom}].

Barkema utilise alors le corpus de Nimègue (130 000 mots), entièrement arboré et qui contient 16 183 syntagmes nominaux relevant de à 1 736 patrons syntaxiques distincts. Il compte le nombre d'occurrences du schéma [{adjectif} {nom}], avec un adjectif « absolu » et un {nom commun singulier} ainsi que le nombre d'occurrences des variantes syntaxiques de ce schéma (dont le passage au pluriel). Il compare alors la fréquence obtenue pour une variation de *cold war* relevant d'un patron donné avec la fréquence attendue. La fréquence attendue d'une telle variation s'obtient en multipliant le nombre total d'occurrences de *cold war* et de ses variations par le nombre de fois où le patron de cette variation se réalise dans les syntagmes libres[24] par rapport au nombre d'occurrences du schéma dont relève *cold war* et de ses variations au sein des syntagmes libres.

Dans les 16 183 syntagmes nominaux du corpus de Nimègue, 1 257 relèvent du schéma [{adjectif absolu} {nom commun singulier}], et 3 171 de ce schéma et de ses variantes syntaxiques. On s'attendrait alors à trouver 49,15 occurrences du schéma de base ((111 + 13) x (1 257 / 3 171)), alors qu'on en trouve 111 : la réalisation au singulier *cold war* est notablement plus fréquente que prévu, ce qui signifie aussi que *cold war* présente moins de variations que le schéma syntaxique dont elle relève ne le permet. L'examen des écarts entre les fréquences attendues et les fréquences observées souligne le fait que la post-modification de *cold war* par un syntagme prépositionnel est moins fréquente qu'on ne s'y attendrait. Il en va de même de la réalisation au pluriel (0 rencontrée, 24,64 occurrences attendues).

3.2.3 Évaluation

L'approche de Barkema pourrait être améliorée. Dans l'idéal, il faudrait pouvoir opérer sur le corpus de Birmingham qui a servi à extraire les variantes de *cold war*. Malheureusement, ce vaste corpus n'est pas muni de structures syntaxiques. Comme Barkema le souligne lui-même, il faudrait pouvoir calculer le poids de chaque réalisation syntaxique d'un schéma fondamental sur le même corpus que celui utilisé pour extraire les variations d'expressions relevant de ce schéma. En effet, rien ne dit que la flexibilité des syntagmes libres ou celle des expressions toutes faites soit la même dans tous les registres. On sait par exemple que l'écrit journalistique contemporain français fait souvent appel à des locutions qui sont détournées : par exemple ce titre de *Libération* du 20 mars 1989

[24] C'est-à-dire ne constituant pas des expressions toutes faites.

après les élections municipales *Coup d'état de grâce* (Fiala et Habert, 1989, p. 91). D'autres registres, comme le discours juridique, sont peut-être plus conservateurs quant à la phraséologie qu'ils véhiculent. Ne disposant pas de corpus arboré de taille suffisante pour pouvoir y observer des phénomènes de flexibilité, Barkema, par la force des choses, en est réduit à « peser » les variations effectives avec une balance réglée sur d'autres données langagières, le corpus de Nimègue, ce qui constitue un biais dont on ne peut pas mesurer les conséquences dans l'immédiat.

Barkema cherche à caractériser la flexibilité du schéma de base dont relève une expression donnée. Une partie des recherches actuelles en syntaxe met l'accent sur les contraintes lexicales gouvernant l'application des règles syntaxiques. Tout adjectif par exemple n'accepte pas la totalité des règles de formation des groupes adjectivaux ni ne rentre dans toutes les places syntaxiques possibles (antéposé / post-posé / après copule). Nous avons vu au chapitre I les restrictions propres aux adjectifs relationnels : construction copulative et adverbe de degré sont impossibles. Les adjectifs de couleur présentent d'autres particularités. Barkema examine simplement les variations du patron [{adjectif absolu} {nom commun singulier}]. C'est sans doute une caractérisation encore trop grossière[25]. Cependant, s'il paraît nécessaire d'utiliser des catégories plus fines, c'est accroître en amont la difficulté de disposer d'un corpus à la fois suffisamment vaste et étiqueté avec suffisamment de finesse.

3.3 La variation de termes en langue de spécialité

Pour obtenir les variations possibles de *cold war*, Barkema utilise un programme qui cherche les phrases comprenant *war* au singulier ou au pluriel et *cold*, pas forcément conjoints ni dans cet ordre. Le tri des séquences effectivement pertinentes est par contre manuel. Dans certaines d'entre elles, *cold* et *war* n'appartiennent pas au même syntagme ou bien ne suivent pas la relation de dépendance présente dans l'expression source.

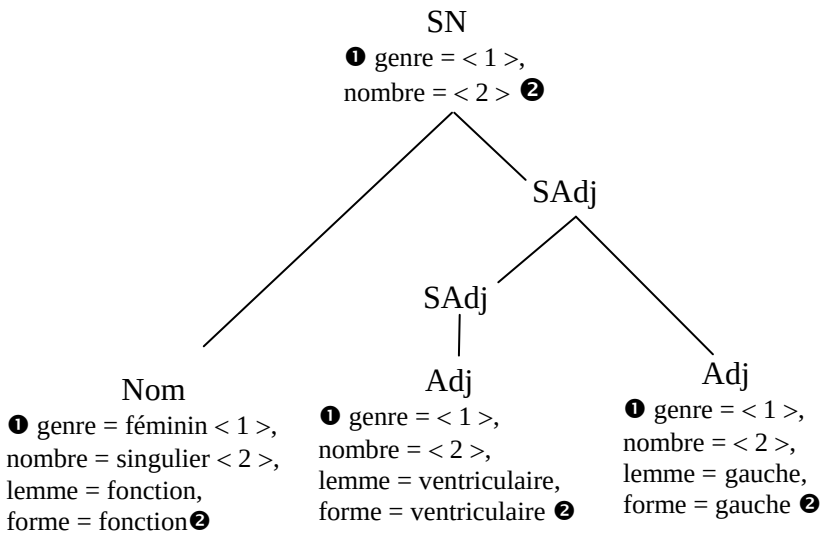
Les recherches de C. Jacquemin (Jacquemin, 1994) sur la variation des termes en langue de spécialité empruntent une démarche radicalement différente où la quête de variations est contrôlée par des connaissances, des règles linguistiques. Au lieu de chercher des séquences en intersection — c'est-à-dire partageant des mots — avec des expressions toutes faites, il s'agit d'engendrer les variations syntaxiques possibles de termes techniques et de vérifier si ces variations se rencontrent effectivement en corpus.

[25] Bien qu'il postule que : « [...] en principe, les expressions libres acceptent l'application de toutes les règles (et sont donc totalement flexibles) » (*ibid.*, p. 44), Barkema montre d'ailleurs quelque inquiétude sur ce point et souhaite vérifier pour des expressions libres comme *the old man ou the bird in the garden* si les variations effectives de ces expressions correspondent bien au profil de variations attendues.

3.3.1 Une représentation syntaxique contrainte des termes

L'objectif est d'inventorier les variations en corpus des termes d'un domaine. On parle aussi de mots-clés ou de *descripteurs* quand ces éléments sont utilisés en informatique documentaire pour indexer des documents. Certains de ces descripteurs sont des mots simples (comme *paradigme* en linguistique). La plupart sont des mots complexes (comme *axe paradigmatique* en linguistique). Ce sont les descripteurs complexes qui sont retenus.

Dans l'optique retenue par Jacquemin, les termes complexes ne sont pas représentés comme des simples suites de mots, mais directement comme des arbres syntaxiques aussi profonds et aussi larges que souhaité. Les relations de dépendance entre les composants sont donc directement indiquées. En outre, les nœuds de ces arbres sont décorés de traits également aussi complexes que nécessaire. Ces nœuds permettent d'assortir les arbres de fines contraintes de bonne formation. Ainsi, pour ***Menelas***, le descripteur *fonction ventriculaire gauche*[26] est représenté de la manière suivante[27] :



La représentation choisie souligne la dépendance de *gauche* par rapport à *ventriculaire* et non à *fonction*. On constate par ailleurs que le nombre de *fonction* est spécifié : ce doit être le singulier, si bien que la séquence *fonctions ventriculaires gauches* ne saurait correspondre à une variation de ce descripteur, puisqu'elle viole l'indication fournie pour le nombre. Les indices entre chevrons indiquent un partage de valeur, ici du genre et du nombre entre la tête et ses modifieurs adjectivaux, ainsi qu'avec les constituants qui les dominent.

[26] L'état fonctionnel du ventricule gauche est crucial en cardiologie. Le ventricule droit ne revêt pas la même importance. *Fonction ventriculaire droite* n'est d'ailleurs pas un mot-clé du domaine.

[27] Dans cet arbre, nous avons laissé comme étiquette du noeud la catégorie du constituant. Nous aurions aussi pu la représenter comme un trait aditionnel : {catégorie=SN...}.

3.3.2 Engendrer des variantes possibles de termes

Une des variations possibles d'un terme de structure [SN Nom [Sadj [Sadj Adj] Adj]] est la modification du syntagme adjectival par un nouvel adjectif à gauche ou à droite. Pour le terme choisi, cela signifie qu'il est *a priori* possible d'en rencontrer la modification suivante :

[SN [Nom fonction][SAdj [Adj x][Sadj [Sadj [Adj ventriculaire]] [Adj gauche]]]

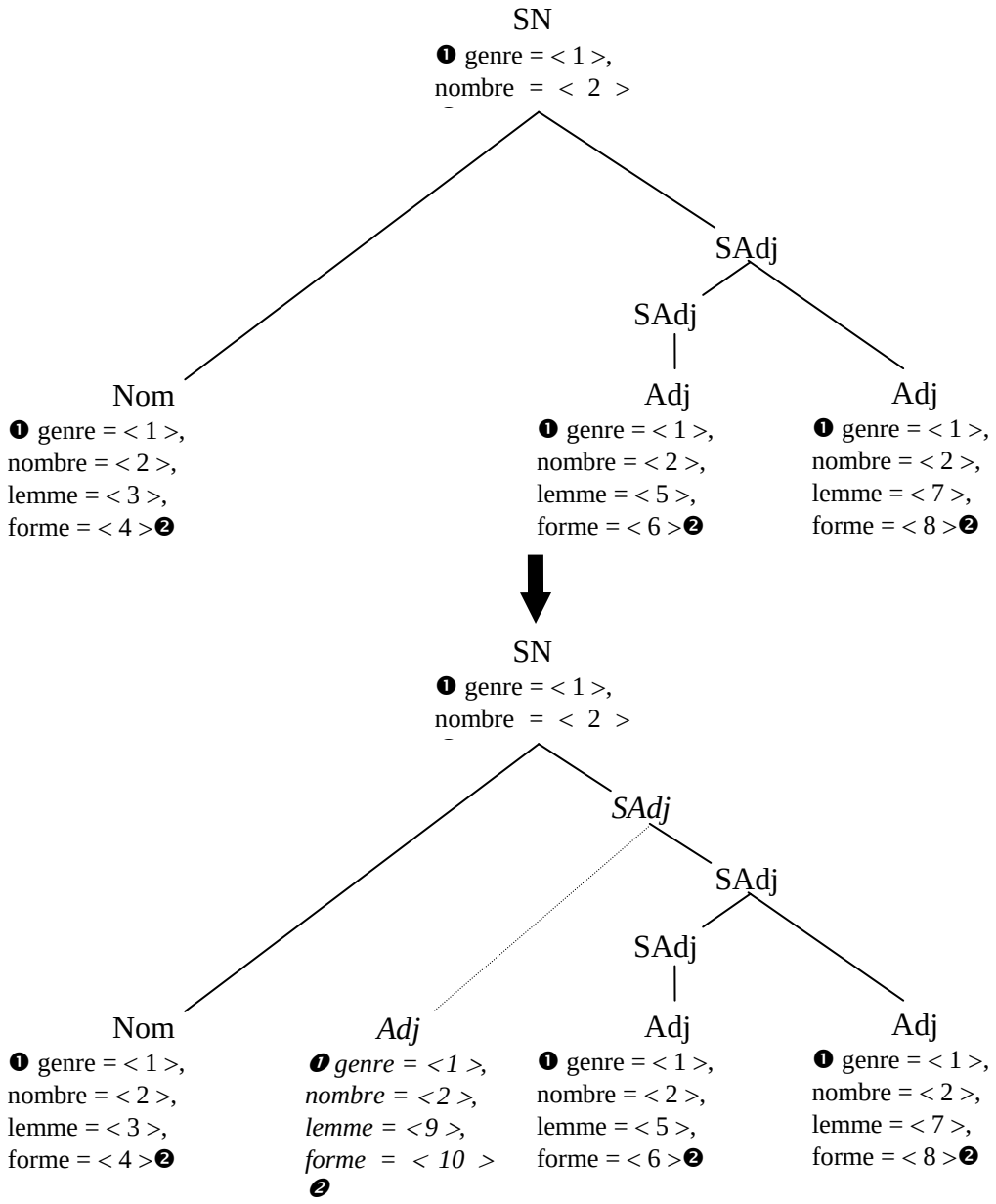
ou bien encore :

[SN [Nom fonction][SAdj [Sadj [Sadj [Adj ventriculaire]] [Adj gauche]][Adj x]]]

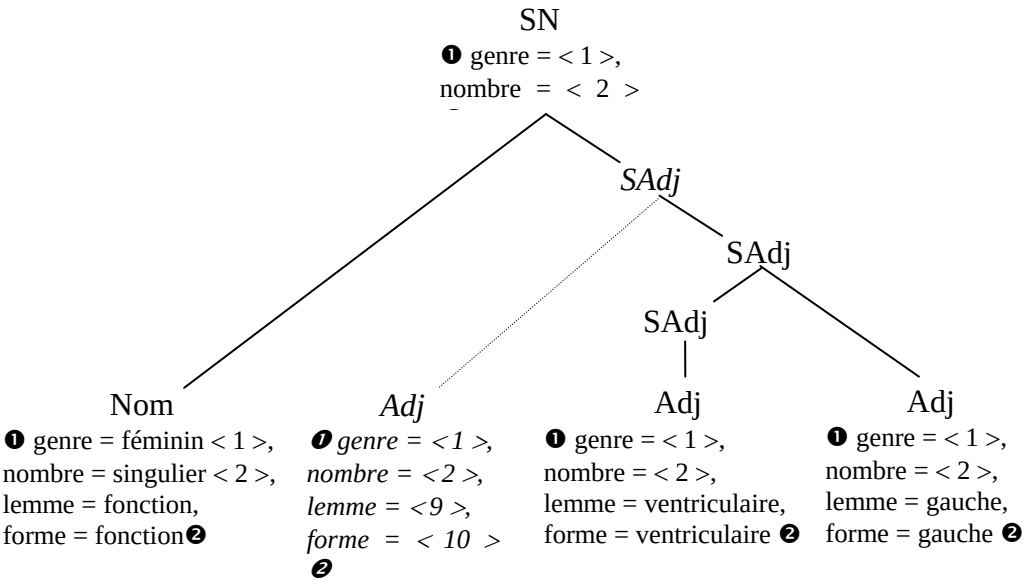
où *x* peut être remplacé par un adjectif quelconque. Les séquences correspondantes sont *fonction* **x** *ventriculaire gauche et fonction ventriculaire gauche* ***x***, dans lesquelles **x** doit être un adjectif.

Des méta-règles servent alors à stipuler les transformations que peuvent éventuellement connaître les descripteurs. Elles prennent en entrée un arbre décrivant un descripteur et produisent en sortie un autre arbre représentant une variation possible de ce descripteur.

La méta-règle suivante :



appliquée à l'arbre représentant le descripteur *fonction ventriculaire gauche* produit l'arbre suivant :



Cet arbre correspond à l'interposition possible d'un adjectif entre *fonction* et *ventriculaire gauche*. Cet adjectif doit s'accorder avec *fonction*. C'est le rôle des indices entre chevrons sur les traits attachés aux noeuds : le trait nombre et le trait genre de l'adjectif inséré doivent avoir la même valeur que les traits correspondants attachés à *fonction*. Le lemme de l'adjectif ajouté n'est pas précisé par contre.

Les méta-règles comprennent donc des « décorations » sur les nœuds. Ces informations permettent de contraindre leur application. On pourrait ajouter par exemple le trait {type = relationnel $\vee$ qualificatif/relationnel} pour empêcher l'engendrement d'une variation avec un adjectif qualificatif : **fonction satisfaisante ventriculaire gauche*. L'adjectif *satisfaisante* portant le trait {type=qualificatif}, il y aurait conflit entre la valeur du trait dans la méta-règle et celle de *satisfaisante*.

Une autre méta-règle peut faire fond sur la valeur du trait nom-base, associé à *ventriculaire* pour engendrer l'arbre correspondant à *fonction du ventricule gauche*, où l'adjectif relationnel *ventriculaire* est remplacé par le syntagme prépositionnel équivalent. Cette transformation peut opérer dans l'autre sens, ce qui permet d'obtenir *infarctus myocardique* à partir d'*infarctus du myocarde*. Ces transformations sont donc conditionnées par la présence de certains traits. Le terme *infarctus du myocarde* peut être transformé en *infarctus myocardique* parce qu'est associé au nœud correspondant à *myocarde* le trait {adjectif-relationnel=myocardique}. Le terme *angine de poitrine* ne pourra pas être transformé de la même manière : l'adjectif *poitrinaire* a le sens d'« atteint de tuberculose poitrinaire » et n'est pas l'adjectif relationnel qui serait nécessaire pour le déclenchement de cette méta-règle[28].

Une méta-règle peut, dans des conditions bien définies, s'appliquer sur les résultats d'autres méta-règles. Les deux méta-règles vues précédemment peuvent par exemple se combiner pour engendrer la variation potentielle *fonction* {adjectif} *du ventricule gauche*.

[28] Pour ajouter ces contraintes, on associe à poitrine le trait {adjectif-relationnel=sans} et à poitrinaire le trait {nom-base=sans}, par exemple.

C. Jacquemin a mis au point par expérimentation sur différents corpus les méta-règles nécessaires pour rendre compte des transformations effectivement rencontrées pour les termes techniques de plusieurs corpus techniques (médecine, métallurgie ...). Toutes les variantes potentielles prévues par les méta-règles et leurs combinaisons à partir d'un ensemble de descripteurs du domaine sont engendrées.

3.3.3 Repérage des variations syntaxiques engendrées

L'analyseur robuste FASTER, développé par C. Jacquemin, recherche ces variations dans un corpus du domaine le plus souvent étiqueté au préalable. C'est un analyseur très particulier : il se cantonne à un type de composant syntaxique, le groupe nominal, et s'en tient aux groupes qui comprennent certaines entrées lexicales, dans des relations de dépendance bien définies et obéissant à des contraintes fines grâce aux traits décorant les nœuds non-terminaux. Dans ***Menelas***, les méta-règles appliquées à *fonction ventriculaire gauche* permettraient de repérer *fonction systolique ventriculaire gauche, fonction ventriculaire gauche systolique* [29], ainsi que *(évalution de la) fonction globale du ventricule gauche* et *fonction du ventricule gauche*. Les transformations non « prévues » aboutiraient à un silence, c'est-à-dire à la non-extraction d'une variation effective. C'est le cas de l'acronyme, attesté : *FVG*. C'est le cas encore du remplacement de la tête par un hyponyme *: cinétique ventriculaire gauche* ou par une périphrase : *état fonctionnel du ventricule gauche*.

3.3.4 Vers une grammaire de la variation terminologique

C. Jacquemin distingue au sein des variations possibles les modifications (la tête ou un dépendant reçoit un modifieur : *fonction systolique ventriculaire gauche*), les permutations (*fonction ventriculaire gauche / fonction du ventricule gauche*) et les coordinations (comme l'hypothétique °*fonction ventriculaire gauche et droite*). Le tri des variations rapportées par l'analyseur entre variantes effectives et « bruit », séquences non reliées au terme de départ, manifeste une dissymétrie de ces trois opérations. La coordination, avec ses contraintes sémantiques, débouche souvent sur des variantes non ambiguës. La modification isole des séquences au statut plus incertain. La permutation enfin aboutit à un taux de bruit encore plus important : il tient au rôle sémantique flou des prépositions dites incolores, en français comme en anglais (*de*, *à*, *of*).

Ce sont là les premiers éléments d'une véritable grammaire de la variation terminologique, capable de caractériser précisément les opérations possibles et leur domaine d'application. On peut même se demander si, à côté de mécanismes très généraux intervenant dans les

[29] Phénomène d'incertitude positionnelle assez fréquent dans ce domaine. En voici un autre exemple : *syndrome douloureux thoracique* / *syndrome thoracique douloureux*.

différents langages spécialisés, ne peuvent pas se rencontrer des régularités particulières à tel ou tel domaine. Dans l'immédiat, cependant, il y a peu de différences d'un corpus à l'autre sur les types de méta-règles à utiliser, ce qui pourrait plaider pour une certaine stabilité de la langue technique au regard des mécanismes syntaxiques employés.

3.4 La recherche de candidats termes

Les deux approches que nous venons de présenter cherchent les variations d'expressions toutes faites de la langue générale ou de termes de langues de spécialité. On part donc de séquences répertoriées dont on cherche en corpus des réalisations modifiées. Le travail que nous examinons maintenant est orienté par l'objectif complémentaire, l'acquisition terminologique, c'est-à-dire repérer les termes d'un domaine quelconque qui n'ont pas encore été répertoriés. Il s'insère dans un contexte industriel, la Direction des Etudes et Recherches d'Electricité de France (DER-EDF).

Une grande entreprise industrielle comme EDF doit maîtriser des flux d'informations électroniques immenses : rapports de recherche internes, articles et publications glanées sur les réseaux, documents destinés au public, etc. Il importe de pouvoir rapidement retrouver l'information pertinente dans cette masse de données, par exemple extraire les documents qui parlent d'une notion donnée.

Pour certains domaines, une terminologie a été établie par des documentalistes ou des terminologues. Elle répertorie les principales notions du domaine et leurs réalisations linguistiques : les termes correspondants. Elle comprend éventuellement des liens de synonymie, d'antonymie, d'hyperonymie. Par exemple, on trouvera dans la terminologie du domaine du TALN des termes comme *analyseur syntaxique*, *formalismes d'unification*, *chaînes de Markov*, un lien de synonymie entre *analyseur syntaxique* et *parseur*, un lien d'hyperonymie entre *parseur* et *analyseur robuste* (un *analyseur robuste* est un type de *parseur*). Ces liens sont utilisés pour élargir les recherches effectuées : un système de recherche d'information pourra, grâce à cette terminologie, rapatrier les textes parlant de *parseur* et d'*analyseur robuste* si la demande porte sur les *analyseurs syntaxiques*.

Dans d'autres domaines, il n'y a pas de terminologie disponible. Cette absence peut tenir au coût de la constitution d'une terminologie par des documentalistes. L'évolution extrêmement rapide de certains secteurs peut aussi contrecarrer le dessein de prendre un « instantané » des termes qui y sont employés : l'image produite a toutes chances d'être déformée. Le vocabulaire de la navigation sur les réseaux (Internet, Web) offre un bon exemple de tels changements incessants. L'acquisition terminologique a de manière générale pour objectif d'isoler les dénominations d'un domaine, pour créer ou compléter une terminologie.

D. Bourigault a développé à la DER-EDF Lexter (Bourigault, 1993), un analyseur destiné à isoler les « candidats-termes » présents dans un

corpus de texte « tout-venant », préalablement étiqueté. Il entend par candidats-termes les syntagmes nominaux qui ont un fonctionnement dénominatif. L'hypothèse fondamentale est qu'un analyseur peut « dégrossir » le travail de repérage des dénominations effectives d'un domaine. Clairement, certaines séquences nominales, parce qu'elles font référence au cotexte ou au contexte, n'ont pas la généricité requise pour des dénominations (Kleiber, 1984). Par exemple *le maintien de sa température* ne serait pas retenu, en raison du possessif, tandis que *le maintien de température*, voire *le maintien de la température* le seraient : le déterminant zéro et le déterminant défini sont compatibles avec une lecture dénominative.

3.4.1 Isoler les groupes d'allure dénominative

La première étape du travail de Lexter consiste à isoler les groupes nominaux d'allure dénominative « maximaux ». L'approche retenue ne s'appuie pas au premier chef sur des règles de structuration du groupe nominal en français. Il s'agit au contraire au départ de repérer les frontières, c'est-à-dire les catégories et suites de catégories qui forment les bornes, exclues, d'un tel constituant. Dans la séquence (*ibid.* p. 108) :

> le circuit d'aspersion de l'enceinte de confinement assure le maintien de sa température nominale de fonctionnement après une augmentation de pression

les éléments *assure*, *de sa*, et *après une* sont considérés comme des frontières. Le verbe est la limite d'un groupe nominal ordinaire. Par contre, *de sa* ne peut servir à articuler deux parties d'une dénomination

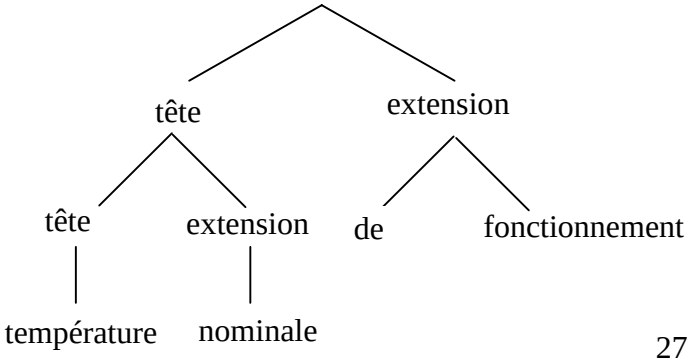
Tête : température nominale

Tête : température

Expansion : nominale

complexe, *après une* non plus. On voit donc se superposer deux types de contraintes : l'une qui cherche à isoler les groupes nominaux, l'autre qui au sein de ce type de constituant, filtre ceux qui peuvent constituer des dénominations. Les groupes retenus sont *: circuit d'aspersion de l'enceinte de confinement*, *maintien, température nominale de fonctionnement*, *augmentation de pression*.

La deuxième étape ne garde que les groupes complexes : *maintien* est laissé de côté à ce stade. Les groupes sont en effet moins ambigus et apportent davantage d'information. Que l'on compare *données* et *base de*



données ou *analyse de données*. La première expression renvoie à l'informatique, la seconde aux statistiques, *données* tout seul potentiellement aux deux. À cette étape, les groupes sont également décomposés de manière récursive selon un schéma dépendanciel en Tête / Expansion[30]. La représentation de *température nominale de fonctionnement* est alors :

L'intérêt de ce type de décomposition, c'est de permettre les regroupements paradigmatiques qui sont si révélateurs en langage spécialisé. Regroupement sur les têtes : on peut mettre à jour des liens de co-hyponymie (entre plusieurs candidats-termes commençant tous par *analyseur* : *analyseur morphologique*, *analyseur syntaxique*, *analyseur robuste*, *analyseur montant* ...) ou d'hyperonymie (entre une séquence courte : *analyseur syntaxique* et une séquence qui la prolonge : *analyseur syntaxique déterministe*). Regroupement sur les expansions : il permet de voir les attributs à « spectre étroit » (qui modifient un nombre restreint de têtes : *déterministe* ne modifie guère qu'*analyseur* en TALN) et ceux qui sont moins spécifiques (*automatique* en informatique ou en TALN).

3.4.2 Le corpus comme norme

Les deux étapes reposent sur un postulat sous-jacent : limiter autant que possible l'appel à un savoir linguistique sur la langue dans son ensemble. Lexter nécessite seulement que le texte analysé ait été étiqueté préalablement pour pouvoir procéder à une analyse syntaxique partielle. Mais Lexter n'utilise ni informations sémantiques ni données de sous-catégorisation : prépositions régies par des noms prédicatifs ou par des adjectifs attendant un régime prépositionnel (*oublieux de*, *attentif à*, etc.). Cet « ascétisme » volontaire s'explique par la conviction, étayée par l'analyse détaillée de textes de domaines techniques distincts, qu'on ne peut pas forcément projeter les connaissances linguistiques générales sur les textes techniques, ou qu'inversement, les textes d'un domaine donné peuvent posséder des particularités combinatoires (des régimes de noms ou d'adjectifs qui le caractérisent) distinctes de celles d'un autre domaine.

Le corrélat logique de ce minimalisme est l'appel à l'apprentissage endogène. C'est considérer le corpus comme sa propre norme, et utiliser les régularités qu'il manifeste pour effectuer découpages et structuration. Lexter est souvent confronté à des ambiguïtés structurelles. Dans la séquence de ***Menelas*** *angine de poitrine instable*, faut-il rattacher *instable* à *poitrine* ou à *angine* ? Un locuteur français doit faire un effort pour simplement percevoir la difficulté. Un étranger qui ne connaîtrait que les mots isolés et pas le terme médical *angine de poitrine* partagerait pourtant l'hésitation de l'analyseur. Lexter dans un premier temps propose les deux découpages pour cette séquence : *[angine de poitrine] instable* et *angine de [poitrine instable]*. Le programme effectue un seul découpage pour les séquences non ambiguës. Dans un deuxième temps, Lexter regarde si l'un des sous-groupes des séquences ambiguës constitue un groupe non ambigu relevé au cours du premier temps. C'est ainsi qu'on rencontre

[30] Ce terme générique recouvre, comme *dépendant*, *modifieur* et *argument*.

dans ***Menelas*** *angine de poitrine*, mais pas *poitrine instable*. On choisit alors le découpage qui contient le groupe non ambigu, ici *[angine de poitrine] instable*. L'évaluation empirique de cette méthode sur différents corpus (*ibid.*, p. 113-114) donne les résultats suivants : dans 75 % des cas, la désambiguïsation obtenue est correcte ; 20 % des séquences restent non désambiguïsées ; 5 % des séquences sont désambiguïsées de manière erronée. Une comparaison de cette approche par apprentissage et d'une résolution des ambiguïtés par des règles *a priori* (Habert *et al.*, 1997) semble donner l'avantage à la première méthode. La délimitation des groupes maximaux repose également partiellement sur l'apprentissage. Certaines séquences constituent en effet des « frontières élastiques », c'est-à-dire qu'elles peuvent tantôt délimiter des groupes nominaux dénominatifs tantôt en faire partie. C'est le cas de *sur* + {article défini} (*ibid.*, p. 109-111). En général, c'est une limite :

1. on raccorde le câble d'alimentation sur le coffret de décharge batterie

Mais ce n'est pas toujours le cas :

2. action sur le bouton poussoir de réarmement
3. action sur le système d'alimentation de secours

En faire une limite intangible, c'est éliminer 2 et 3. L'accepter au sein des candidats-termes conduit à isoler *le câble d'alimentation sur le coffret de décharge batterie*, qui ne constitue certainement pas une séquence dénominative. La solution réside là encore dans l'apprentissage endogène. Il porte cette fois-ci sur les noms suivis d'une séquence sur + {article défini} + contexte droit immédiat. Un premier passage sur le texte relève tous ces contextes. Un second les trie et répartit les noms en deux groupes : ceux qui sont « productifs » avec *sur* (qui figurent dans le texte avec un nombre suffisant d'expansions différentes introduites par *sur* + {article défini}) et ceux qui ne sont suivis qu'exceptionnellement par *sur* + {article défini}. Lexter considère que l'expansion des premiers peut être introduite par *sur* + {article défini} et garde alors les séquences ayant pour tête à un niveau quelconque ces noms suivis d'une expansion introduite par *sur* + {article défini}. Dans les autres cas, *sur* + {article défini} continue à constituer une frontière. L'apprentissage porte donc ici sur des formes de sous-catégorisation.

3.4.3 Vers une grammaire des dénominations complexes possibles

L'ensemble retenu par Lexter est encore nettement trop vaste par rapport à ce qu'un expert du domaine considèrerait comme termes effectifs. Toutefois, l'objectif visé n'est certainement pas une automatisation totale de la mise en évidence des termes d'un domaine. Pour deux raisons fondamentales. La première, c'est que l'utilisation de Lexter sur des corpus variés, de domaines distincts, montre que les règles de bonne formation de termes possibles ne sont pas forcément les mêmes d'un domaine à l'autre. C'est pourquoi l'apprentissage endogène est justement

incontournable. La seconde raison tient à la complexité des mécanismes par lesquels une communauté langagière sélectionne, parmi les dénominations possibles, celles qui deviennent des dénominations effectives. Si l'on peut espérer diminuer la taille de l'ensemble des candidats-termes extraits d'un corpus, repérer ceux d'entre eux qui fonctionnent réellement comme des termes semble difficilement automatisable.

Lexter matérialise, par les séquences qu'il considère comme des bornes, un certain nombre d'hypothèses sur ce qui ne peut pas figurer dans une séquence nominale pour qu'elle puisse être employée comme une dénomination. La démarche suivie dans la mise au point et le test du logiciel sur des corpus variés ont conduit à rajouter d'autres règles, également négatives. La démarche est proche de celle utilisé pour l'étiquetage (cf. chapitre VIII) : peu à peu, on dégage les régularités à l'œuvre et on met au point des procédures qui s'appuient sur elles. Au total, Lexter, au delà des procédures mises en œuvre, essaie donc de formaliser partiellement la notion de dénomination possible.

3.5 Enjeux pratiques et théoriques

3.5.1 Améliorer la description lexicographique

Barkema (*ibid.*) souligne que le degré de flexibilité d'une expression est rarement indiqué par les dictionnaires qui donnent cette séquence. Le dictionnaire *COBUILD* (Sinclair *et al.*, 1987) fait partiellement exception : pour *moment of truth* (*minute de vérité*), est ainsi indiqué que la seule modification possible de l'expression est l'utilisation au pluriel. Le repérage des réalisations possibles tel qu'il est effectué par Barkema permet d'enrichir la description lexicographique des expressions concernées. Il en va de même en terminologie spécialisée, où les résultats de FASTER isolent des variantes à intégrer dans les ressources lexicales.

En acquisition terminologique, Lexter permet d'enrichir le répertoire des termes à utiliser. L'écrémage par ce programme des dénominations possibles facilite le travail du lexicographe spécialisé. Les concordances de mots fréquents sont en effet souvent très difficiles à dépouiller et à organiser. Le découpage opéré en tête / expansion et les regroupements par têtes et par expansions offrent au contraire une vision synthétique du fonctionnement syntagmatique et paradigmatique des noms « pivots » du texte étudié. L'un des résultats de Lexter est d'ailleurs un réseau terminologique hypertextuel. Chaque candidat-terme est relié à sa tête et à son expansion et d'autre part à tous les candidats-termes dont il est lui-même tête ou expansion. Le lien aux documents de départ permet de replonger les séquences extraites dans leur contexte. Le tout permet à un connaisseur du domaine de séparer dans de bonnes conditions les termes effectifs des groupes parasites.

L'acquisition terminologique, possible avec FASTER, réalisée avec LEXTER, est une tâche dont les résultats sont difficiles à évaluer

objectivement. Il n'existe pas de corpus de tests où les termes pertinents seraient isolés et qui serviraient ainsi d'aune pour mesurer l'apport de ces outils. En outre, le projet de créer de tels corpus est peut-être chimérique. Ce sont des ensembles de termes distincts qui risquent d'être repérés par des experts différents en fonction de leurs préoccupations et de leurs points de vue. Un spécialiste d'épidémiologie et un cardiologue n'identifieront pas forcément les mêmes séquences dans ***Menelas***.

3.5.2 Distinguer variantes et variations

Dans les recherches de Barkema comme dans celles de C. Jacquemin, une fois repérées des variations autour de séquences de départ, termes ou expressions toutes faites, une des difficultés consiste à isoler les variantes effectives, celles qui fonctionnent comme des réalisations possibles pour les expressions considérées.

En langue de spécialité, c'est le recours à un expert qui seul permet de trancher. En langue générale, il faut éliminer les variations qui constituent des défigements intentionnels, des jeux de langage[31], et non des variantes des expressions de départ. Ainsi, les exemples 7 (*a not-so-cold war against Kaddafi*) et *9 (a period of cold and hot civil war which ended with Hitler's invasion of Austria*) de Barkema semblent relativement éloignés du sens originel, qui renvoie au monde d'après Yalta et qu'évoquent les exemples 5, 6 et 11 par exemple.

3.5.3 Importance quantitative de la variation

C. Jacquemin a évalué les résultats de l'extraction de variations de descripteurs engendrées par méta-règles. Il a utilisé à l'INIST[32] un corpus de 125 000 mots dans le domaine de la physique de la métallurgie et un sous-ensemble du lexique terminologique PASCAL utilisé à l'INIST pour l'indexation manuelle (6 621 termes liés à la physique et à la chimie de la métallurgie). Les méta-règles étaient au nombre de 112. Les occurrences de termes et de leurs variantes couvrent (en nombre de mots) 7 % de la surface du corpus, les variantes représentant 28 % de cette « zone terminologique ». Les variantes validées représentent 15 % des occurrences de termes. Cette estimation semble d'ailleurs une valeur plancher, au regard d'expériences sur d'autres langues et d'autres corpus. La variation terminologique est donc loin d'être négligeable, contrairement à un préjugé répandu : les termes seraient les « noms » univoques et stables des notions d'un domaine. Les résultats de Barkema vont dans le même sens, cette fois-ci pour la langue générale. Il semble en effet qu'au total, l'intuition linguistique ou, en langage spécialisé, celle d'un terminologue voire d'un expert du domaine, sous-estime les variantes effectives des dénominations complexes. Le recours au corpus renouvelle

[31] Cf. (Authier-Revuz, 1995).
[32] Institut National pour l'Information Scientifique et Technique - CNRS.

donc l'analyse de la variation de ces unités polylexicales.

3.5.4 Caractériser la flexibilité « normale »

Barkema distingue (*ibid.*, p. 40-41 ; 1993) trois dimensions qui s'articulent : la flexibilité syntaxique : la possibilité pour un groupe de se voir appliquer tout ou partie des règles du constituant dont il relève, la compositionnalité : le fait que le sens de la séquence soit ou non fonction du sens de ses constituants, et enfin la « collocativité » : les préférences d'emploi d'un mot (comme dans l'association privilégiée *économiste* et *distingué* où l'adjectif peut être modifié, coordonné, etc.). Les travaux sur le « figement » ont sans doute eu tendance à confondre ces dimensions qui sont partiellement indépendantes. Le recours aux corpus permet de cerner précisément la première d'entre elles.

Barkema montre comment un corpus arboré permet de fournir une caractérisation fine de la flexibilité attendue pour un schéma syntaxique donné. On peut alors porter un jugement sur les réalisations effectives d'une expression relevant de ce schéma. L'emploi d'un corpus arboré souligne le fait que certaines réalisations d'un schéma syntaxique sont plus probables que d'autres, pondérations qui échappent pour l'essentiel à la conscience d'un locuteur. Les contraintes sur la flexibilité ont suscité depuis longtemps les recherches. Barkema essaie de caractériser précisément l'autre pôle de l'opposition : la flexibilité « normale ». C'est effectivement une tâche nécessaire pour pouvoir parler en connaissance de cause de degré de figement. Le corpus offre le moyen de pondérer les règles applicables à un constituant donné.

4. UTILISER DES PARSEURS ET DES CORPUS ARBORES

4.1 Utiliser des parseurs

La mise au point des parseurs nécessite des mécanismes complexes qui sont dans l'immédiat plutôt l'apanage d'informaticiens que de linguistes. L'écriture et l'ajustement de grammaires pour des analyseurs robustes nécessite par exemple des mécanismes de pistage (de trace, disent les informaticiens) : examiner en détail le processus même d'analyse d'une phrase, pour vérifier la pertinence des règles employées, ajouter des règles nécessaires, etc., comme dans le « banc d'essai » de grammaires de l'université de Nimègue (Nederhof et Koster, 1993, p. 174). Ou encore un générateur qui produit aléatoirement des phrases en fonction des règles et du lexique utilisés : cela permet de repérer certains incohérences ou le laxisme sur certains points de la grammaire.

L'utilisation de parseurs pour la constitution de corpus arborés suppose encore dans une coopération étroite entre linguistes et informaticiens. Les

exemples de telles coopérations sont encore rares : le groupe de Nimègue***, Lancaster Treebank*** (Black *et al.,* 1993) et ***Penn Treebank*** (Marcus *et al.,* 1993).

4.2 Utiliser des corpus arborés

Pour parler des corpus annotés syntaxiquement, on utilise également les dénominations de banques d'arbres (*treebank*) et de bases de données syntaxiques (*syntactic database*) (Souter et Atwell, 1994, p. 142). Ces appellations pourraient faire croire à une utilisation aisée des corpus arborés, au même titre que les bases de données du commerce. Il n'en est rien.

Au sens informatique, une base de données associe des tables d'information représentant des relations dans un sens assez proche de celui de la théorie des ensembles et des méthodes pour exprimer des requêtes sur les informations présentes dans une collection de tables, ces méthodes faisant appel à l'algèbre relationnelle qui permet d'exprimer ces requêtes sans entrer dans les détails de la mise en œuvre des opérations. Dans une « base de données syntaxiques *»,* il y a bien accumulation d'informations (et un certain « démembrement », puisque les analyses sont simplement juxtaposées). Mais n'y sont présents ni une formalisation générale des données présentes (on a déjà souligné l'éclatement des pratiques d'annotation syntaxique[33]) ni un langage de requête adéquat, ni même la possibilité d'ajouter ou de retirer des informations, ce que permettent les bases de données. La variété des informations présentes et leur structuration complexe (en termes d'enchâssement de constituants mais aussi de liens horizontaux — par exemple pour les co-références ou pour les discontinuités — ou encore de structures de traits décorant les nœuds) constituent, il est vrai, un défi à la formalisation.

C'est LDB (*Linguistic DataBase*) qui se rapproche le plus d'un outil de gestion et d'interrogation de vastes ensembles de phrases arborées. Cet outil a d'ailleurs été utilisé pour d'autres ensembles arborés que ceux de l'université de Nimègue pour lesquels il a été conçu[34]. Il est possible donc de transformer un corpus arboré pour le rendre interrogeable par LDB[35]. Halteren et Heuvel (1990) offrent une présentation approfondie de l'ensemble des manipulations offertes. Les interrogations peuvent associer les contraintes structurelles (un nœud de telle catégorie dans telle position de dominance ou de dépendance par rapport à tel autre nœud) et des conditions sur les « décorations *»* des nœuds qui peuvent comporter un certain nombre d'étiquettes (ce qui équivaut à un système de traits). On peut par exemple chercher les phrases à constructions bi-transitives (du type *I gave him a book*), ou encore construire un tableau

[33] (Souter, 1993) le montre en détail sur 7 collections ou corpus arborés.
[34] Par exemple, pour la version parsée, au sein de l'équipe ASCOT de l'université d'Amsterdam, du *Longman Dictionary of Contemporary English* (*LDOCE*) (Souter, 1993, p. 204).
[35] Voire d'autres données arborescentes, comme des définitions de dictionnaire (Halteren et Heuvel, 1990, p. 10).

indiquant le nombre de noms modifiés par un groupe adjectival préposé et leurs correspondants avec adjectif postposé, et le nombre de noms non modifiés. C'est LDB que Barkema a utilisé pour déterminer les différentes réalisations syntaxiques du patron de base adjectif nom singulier.

Comme pour l'étiquetage, deux grandes fonctionnalités sont nécessaires. Elles doivent d'ailleurs pouvoir se combiner. D'abord filtrer les arbres répondant à des contraintes arbitrairement complexes. Les outils actuellement disponibles (comme ceux fournis avec ***Penn Treebank***) sont encore rudimentaires et en tout état de cause non génériques : ils sont faits pour traiter d'arbres selon un format d'encodage donné et ne travaillent pas à un niveau de généralité suffisant. Deuxième fonctionnalité : transformer des arbres. Il peut s'agir de changer des étiquettes pour faciliter l'interprétation, ou de restructurer des sous-arbres. Alors que les techniques de transduction d'arbre sont bien maîtrisées en informatique, leur mise à la disposition des utilisateurs de corpus arborés reste pour l'essentiel à réaliser[36].

[36] Cf. (Habert *et al.*, 1997) pour une utilisation de la transduction d'arbres pour la comparaison de deux outils d'acquisition terminologique.