LIGIA CAFÉ

LA DESCRIPTION ET L'ANALYSE DES UNITÉS TERMINOLOGIQUES COMPLEXES EN LANGUE PORTUGAISE (VARIÉTÉ BRÉSILIENNE) : UNE CONTRIBUTION À L'AUTOMATISATION DE LA BANQUE DE DONNÉES TERMINOLOGIQUES DU BRÉSIL (BRASILTERM)

Tome I

Thèse
présentée
à la Faculté des études supérieures
de l'Université Laval
pour l'obtention
du grade de Philosophiae Doctor (Ph. D.)

Département de langues et linguistique
FACULTÉ DES LETTRES
UNIVERSITÉ LAVAL
QUÉBEC

AVRIL 1999

© Ligia Café, 1999

Résumé 1

Notre recherche vise à contribuer à la conception d'un logiciel d'extraction des unités terminologiques complexes (UTC) contenues dans des textes en langue portugaise (variété brésilienne). Pour ce faire, nous avons choisi comme corpus de données les UTC contenues dans des textes écrits en portugais sur la biotechnologie des végétaux. Ces UTC sont décrites et analysées selon leurs aspects morphologiques, syntaxiques, sémantiques et pragmatiques à la lumière de la théorie de la grammaire fonctionnelle de Simon Dik. L'interprétation de ces aspects résulte en une connaissance plus approfondie des mécanismes de fonctionnement des UTC, ce qui sert également de base pour établir des pistes pour l'automatisation.

Ligia Café
étudiante

Pierre Auger
directeur

Enilde Faulstich
co-directrice

Résumé 2

L'analyse du comportement linguistique des unités terminologiques complexes (UTC) est une étape primordiale pour la conception d'un logiciel d'extraction de terminologie en contexte. Toutefois, c'est loin d'être une tâche facile à accomplir. Sur le plan théorique, la littérature qui traite de ce sujet présente différentes méthodes d'analyse qui touchent tant le comportement de l'UTC à l'intérieur de la phrase que les dépendances entre les composants d'une UTC. Dans la pratique, l'informatisation de ces méthodes va des logiciels fondés sur une technique essentiellement statistique jusqu'à ceux qui tirent profit des connaissances fondamentales en linguistique pour concevoir leurs programmes. Notre recherche vise justement à contribuer à la conception de logiciels d'extraction des unités terminologiques complexes qui adoptent une démarche linguistique. Pour ce faire, nous avons choisi comme corpus de données les UTC contenues dans des textes écrits en portugais (variété brésilienne) sur la biotechnologie des végétaux. Ces UTC sont décrites et analysées selon leurs aspects morphologiques, syntaxiques, sémantiques et pragmatiques à la lumière de la théorie de la grammaire fonctionnelle de Simon Dik. L'application de la méthode de la prédication, conçue pour cette théorie, nous a permis de dégager de notre corpus dix règles de formation des UTC. À partir de l'analyse de l'usage des mots de la langue, nous interprétons ces règles pour mieux comprendre le fonctionnement des UTC. Le résultat de cette interprétation linguistico-fonctionnelle est utilisé pour élaborer des stratégies pour la conception d'un logiciel d'extraction terminologique dans un environnement d'intelligence artificielle orienté vers la modélisation des connaissances linguistiques. Nous proposons des pistes aux niveaux morphosyntaxique et sémantique qui donnent accès à une base de connaissances de la prédication. Tout le contenu de cette base est fondé sur les résultats de l'analyse et de l'interprétation des données employées dans notre recherche.

Ligia Café
étudiante

Pierre Auger
directeur

Enilde Faulstich
co-directrice

Avant-propos

C’est au Brésil que l'idée du doctorat m'est venue. Travaillant dans le domaine de la terminologie depuis la fin de ma maîtrise, je me suis rendue compte que mes réflexions ne seraient pas exactes si je ne les fondaient pas sur des bases linguistiques. C'est alors qu'est apparue une personne très spéciale : la professeure Enilde Faulstich. C’est elle qui m’a aimablement tendu la main et qui m’a montré le chemin. Ce chemin m'a conduite à Québec, plus précisément à l’Université Laval. Dans cette université, j’ai trouvé les connaissances et les encouragements nécessaires pour entreprendre mon doctorat. Celui-ci n'aurait pu être possible sans l'aide de plusieurs personnes.

Mes premiers remerciements s’adressent à mon directeur de thèse M. Pierre Auger. Sa grande disponibilité et son ouverture d’esprit pour guider une recherche sur une langue étrangère pour lui m’ont beaucoup touchée.

J’aimerais également témoigner ma reconnaissance à ma co-directrice Mme Enilde Faulstich. Sans ses précieux conseils et sa profonde connaissance du sujet, je pense que je n'aurais pas pu parvenir à l’étape finale. Il faut aussi mentionner que sa supervision minutieuse a été ponctuée de moments d’amitié inoubliables.

J’aimerais adresser des remerciements à la Capes, l’organisme subventionnaire brésilien, et à l’Ibict, l’institut auquel je suis rattachée au Brésil, qui par leur appui financier m’ont permis de me consacrer exclusivement à la recherche.

Je désire également exprimer mes remerciements à mon pré-lecteur M. Jean-Claude Boulanger dont les réflexions toujours à propos ont la noble qualité de ne pas offusquer la créativité de ceux qui font leurs premiers pas comme chercheurs.

À mes amis, je pense que mille mots n’auraient pas suffit pour finalement vous dire *merci de tout cœur* ! Les énumérer ne serait pas prudent, mais je le fais quand même en

espérant que ceux qui ne sont pas dans la liste me pardonneront... Un gros merci à Tina, à Ludo, à Judith, à Dorothy, à Claude Rocheleau, à mes amies brésiliennes Marcela, Marisa et Helena, à mon amie galicienne Iolanda, à M. Perestrelo, ainsi qu'à M. Lebel et à son épouse Monique. Un merci spécial à mon amie Guylaine pour son talent et sa patience pour corriger mon français, mais surtout pour son soutien continu dans les moments de crise ! Je tire mon chapeau à cette fille au grand cœur !

Mes parents, je vous aime si fort ! Vous êtes toujours là malgré tous les kilomètres qui nous séparent. C'est grâce à vous que j'ai réussi ! Je dis aussi merci à tous les autres membres de la famille, plus spécialement à mes beaux-parents, qui ont suivi avec intérêt les différentes étapes de la recherche.

À Marcos, je le remercie infiniment d'avoir accepté de participer à mon aventure doctorale, avec mes hauts et mes bas, et de m'avoir accordé son support moral et technique. Finalement, je voudrais remercier les deux petites étoiles de mon cœur, Flávio et Gustavo, qui ont soulagé plusieurs fois mes inquiétudes avec leurs yeux brillants et leurs sourires enchantés. C'est à eux deux que je dédie ma thèse.

Table des matières

Résumé 1 _____ i
Résumé 2 _____ ii
Avant-propos _____ iii
Table des matières _____ v
Liste des figures _____ viii
Liste des tableaux _____ x

0. Introduction _____ 1

1. Cadre théorique _____ 5

1.1 Introduction _____ 5

1.2 Linguistique fonctionnelle _____ 5

1.3 Théorie de la grammaire fonctionnelle de Dik _____ 9
1.3.1 Bases de la théorie _____ 10
1.3.2 Méthode de la prédication _____ 12
1.3.2.1 Structure de la prédication _____ 13
1.3.2.2 Différences entre arguments et satellites _____ 15
1.3.2.3 Fonctions de la prédication _____ 17

1.4 Application de la théorie de la grammaire fonctionnelle de Dik _____ 23
1.4.1 Adaptations de base _____ 24
1.4.1.1 Règle de formation _____ 25
1.4.1.2 Fonctions sémantiques, syntaxiques et pragmatiques _____ 29

1.5 Conclusion _____ 40

2. État de la question _____ 42

2.1 Introduction _____ 42

2.2 Fonctionnalisme dans les études terminologiques _____ 43
2.2.1 Description et analyse des unités terminologiques complexes _____ 44
2.2.2 Critères de figement _____ 55

2.3 Informatique et terminologie _____ 61
2.3.1 Extraction automatisée d'unités terminologiques complexes _____ 66
2.3.2 Traitement et récupération terminologique _____ 81
2.3.3 Base de connaissances terminologiques _____ 85

2.4 Problématique de la recherche _____ 90

2.5 Objectif de la recherche _____ 92

2.6 Conclusion 92
3. Constitution et informatisation du corpus 94
3.1 Introduction 94
3.2 Choix du domaine 94
3.3 Typologie des textes 98
3.4 Informatisation du corpus 101
3.5 Critères de validation de la terminologie recueillie 105
3.6 Systématisation des données 108
3.6.1 Élaboration de la base *Access* 108
3.6.1.1 Description des tables 110
3.6.1.2 Formulaire d'entrée des données 113
3.6.1.3 Types de sorties générées 116
3.7 Conclusion 118
4. Analyse et interprétation des données 119
4.1 Introduction 119
4.2 Règles de formation 119
4.2.1 [ϕ [x_1]] 125
4.2.1.1 Base 125
4.2.1.2 Argument 135
4.2.1.3 Analyse morphosyntaxique 145
4.2.1.4 Relation conceptuelle entre la base et les arguments 152
4.2.1.5 Synthèse de la règle [ϕ [x_1]] 160
4.2.2 [[ϕ [x_1]] [y_1]] 161
4.2.2.1 Interprétation fonctionnelle de la prédication étendue : le satellite 163
4.2.2.2 Analyse morphosyntaxique 166
4.2.2.3 Relation conceptuelle entre la base, l'argument et le satellite 168
4.2.3 [ϕ [y_1]] 170
4.2.3.1 Satellite 171
4.2.3.2 Analyse morphosyntaxique 172
4.2.3.3 Relation conceptuelle entre la base et le satellite 174
4.2.4 [ϕ [x_1 [$x_{1/1}$]]] 177
4.2.4.1 Argument et expansion 177
4.2.4.2 Argument-adverbe et expansion 179
4.2.4.3 Analyse morphosyntaxique 183
4.2.4.4 Rôle de l'argument-expansion dans les relations conceptuelles 185
4.2.5 [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 187
4.2.5.1 Niveaux de rapports entre argument et expansions 188
4.2.5.2 Rôle syntaxique du complément circonstanciel 190
4.2.5.3 Analyse morphosyntaxique 193
4.2.5.4 Rôle de l'argument-expansion dans les relations conceptuelles 194

4.2.6 [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] 199
4.2.6.1 Base et arguments 199
4.2.6.2 Analyse morphosyntaxique 202
4.2.6.3 Relations conceptuelles entre la base et ses arguments 205
4.2.7 Cas particuliers 207
4.2.7.1 [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]] 208
4.2.7.2 [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]] 213
4.2.7.3 [ϕ [x_1] [x_2 [$x_{1/2}$]]] 218
4.3 Conclusion 224
5. Stratégies pour l'automatisation de l'extraction des UTC 228
5.1 Introduction 228
5.1.1 Considérations de base 229
5.1.1.1 Pistes au niveau morphosyntaxique 232
5.1.1.2 Pistes au niveau sémantique 236
5.2 Conclusion 241
6. Conclusion 243

Bibliographie 248
Annexes 260
Annexe 1 — Description des UTC
Annexe 2 — Fonctions sémantiques
Annexe 3 — Texte utilisé pour attester les pistes de notre logiciel
Index des auteurs cités

Liste des figures

Figure 1.1 — Structure générale de la prédication 13
Figure 1.2 — Paramètres pour l'établissement des fonctions sémantiques 18
Figure 1.3 — Schéma des interrelations entre les paramètres 20
Figure 1.4 — Hiérarchie de la fonction sémantique 21
Figure 1.5 — Règle de formation des UTC 26
Figure 1.6 — Représentation de la règle et des fonctions sémantiques de l'UTC 28
Figure 3.1 — Classement hiérarchique du domaine de la biotechnologie des végétaux 95
Figure 3.2 — Sélection des termes simples 103
Figure 3.3 — Repérage des termes simples faisant partie d'une UTC 104
Figure 3.4 — UTC suivies de leurs contextes 105
Figure 3.5 — Tables et relations de la base Access 109
Figure 3.6 — Formulaire d'entrée des données 113
Figure 3.7 — Formulaire d'entrée des données présentant la liste des occurrences de la table « type de prédication » 114
Figure 3.8 — Exemple d'interrogation 115
Figure 3.9 — Exemple de données en mode feuille 116
Figure 3.10 — Rapport concernant les UTC 117
Figure 4.1 — Hiérarchie de la fonction sémantique des bases-sujets pour la règle [ϕ [x_1]] 134
Figure 4.2 — Hiérarchie de la fonction sémantique pour les arguments de fonction syntaxique adjoint de nom pour la règle [ϕ [x_1]] 142
Figure 4.3 — Formules syntaxiques et représentation en formules morphologiques de la règle [ϕ [x_1]] 146
Figure 4.4 — Arbres syntagmatiques d'*agregado de célula/agregado celular* 151
Figure 4.5 — Arbres syntagmatiques d'*axila da folha/axila foliar* 151
Figure 4.6 — Locution adverbiale nucléaire dans la structure [ϕ [x_1]] 164
Figure 4.7 — Locution adverbiale étendue dans la structure [[ϕ [x_1]] [y_1]] 165
Figure 4.8 — Formules syntaxiques et représentation en formules morphologiques pour la règle [[ϕ [x_1]] [y_1]] 166
Figure 4.9 — Rapports entre les composants de l'UTC pour la règle [[ϕ [x_1]] [y_1]] 168
Figure 4.10 — Formule syntaxique et représentation en formule morphologique pour la règle [ϕ [y_1]] 173
Figure 4.11 — Dérivation de la règle [ϕ] [y_1]] à partir de la règle [[ϕ [x_1]] [y_1] 176
Figure 4.12 — Rapports selon le type de relation entretenue par l'argument 178
Figure 4.13 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1 [$x_{1/1}$]]] 183
Figure 4.14 — Fonctions syntaxiques pour la règle [ϕ [x_1 [$x_{1/1}$]]] 185
Figure 4.15 — Relation conceptuelle *produit un résultat* pour la règle [ϕ [x_1 [$x_{1/1}$]]] 187
Figure 4.16 — Rapports entre les composants de la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 188
Figure 4.17 — Représentation des rapports entre les composants de la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 190
Figure 4.18 — Représentation des compléments circonstanciels 191
Figure 4.19 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 193
Figure 4.20 — Relation conceptuelle et fonctions sémantiques pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 195
Figure 4.21 — Première relation conceptuelle *produit un effet* pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 196
Figure 4.22 — Seconde relation conceptuelle *produit un effet* pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 197
Figure 4.23 — Relation conceptuelle *résulte de* pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 198
Figure 4.24 — Changement de fonction sémantique dans la relation conceptuelle *résulte de* pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 199
Figure 4.25 — Représentation en arbres syntagmatiques de la règle [ϕ [x_1 [$x_{1/1}$]]] 200
Figure 4.26 — Représentation en arbres syntagmatiques de la règle [ϕ [x_1] [x_2]] 200
Figure 4.27 — Représentation en arbres syntagmatiques de la règle [ϕ [x_1] [x_2] [x_3]] 201

Figure 4.28 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1] [x_2]] 202
Figure 4.29 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1] [x_2] [x_3]] 202
Figure 4.30 — Catégorie grammaticale et fonction syntaxique de *planta progenitora feminina* 204
Figure 4.31 — Catégorie grammaticale et fonction syntaxique de *infecção de plantas por vírus* 204
Figure 4.32 — Relations conceptuelles et fonctions sémantiques de *célula vegetal nutriz* 206
Figure 4.33 — Différentes fonctions sémantiques pour la même base d'une UTC 207
Figure 4.34 — Représentation graphique de l'UTC *bactéria de vida livre fixadora de nitrogênio* 209
Figure 4.35 — Processus de formation de l'UTC *bactéria de vida livre fixadora de nitrogênio* 210
Figure 4.36 — Rapports selon la relation entretenue entre la base, les arguments et les expansions 210
Figure 4.37 — Formule syntaxique et morphologique pour la règle [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]] 211
Figure 4.38 — Alternance de catégories grammaticales selon l'ensemble argument et expansion 211
Figure 4.39 — Formation de SA dans l'UTC *bactéria de vida livre fixadora de nitrogênio* 212
Figure 4.40 — Relations conceptuelles et fonctions sémantiques de *bactéria de vida livre fixadora de nitrogênio* 213
Figure 4.41 — Rapports selon la relation entre la base, les arguments et les expansions pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]] 214
Figure 4.42 — Nouvelle représentation des rapports entre les composants de l'UTC *propagação clonal com fidelidade ao tipo original* 215
Figure 4.43 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]] 215
Figure 4.44 — Arbre syntagmatique de *propagação clonal com fidelidade ao tipo original* 216
Figure 4.45 — Relations conceptuelles et fonctions sémantiques de *com fidelidade ao tipo original* 217
Figure 4.46 — Rapports selon la relation entre la base, les arguments et l'expansion pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]] 219
Figure 4.47 — Rapports selon la relation entre la base, l'argument, l'argument-adverbe et l'expansion pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]] 219
Figure 4.48 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]] 221
Figure 4.49 — Relations conceptuelles et fonctions sémantiques de *célula somática geneticamente diferente* 222
Figure 4.50 — Relations conceptuelles et fonctions sémantiques de *clone sadio livre de virose* 223
Figure 4.51 — Relations conceptuelles et fonctions sémantiques de *célula interna altamente vacuolizada* 224
Figure 5.1 — Structure générale du traitement automatique pour l'extraction des UTC 230

Liste des tableaux

Tableau 1.1 — Fonctions syntaxiques et sémantiques des UTC 33
Tableau 2.1 — Classification des concepts 51
Tableau 2.2 — Principales différences entre groupe nominal ordinaire ou libre et nom composé 56
Tableau 2.3 — Comparaison entre les approches linguistique et statistique 76
Tableau 2.4 — Grille d'évaluation de L'Homme et al 80
Tableau 3.1 — Pourcentage d'UTC trouvées dans les ouvrages de référence 107
Tableau 3.2 — Table « éléments » et ses champs 111
Tableau 3.3 — Table « utc » et ses champs 112
Tableau 4.1 — Pourcentage d'UTC par règle de formation 121
Tableau 4.2 — Pourcentage d'UTC par type de base pour la règle [ϕ [x_1]] 126
Tableau 4.3 — Pourcentage d'UTC de base active par suffixe pour la règle [ϕ [x_1]] 129
Tableau 4.4 — Suffixes et fonctions sémantiques des bases actives pour la règle [ϕ [x_1]] 130
Tableau 4.5 — Terminologie employée selon le type de base pour la règle [ϕ [x_1]] 131
Tableau 4.6 — Notion sémantique employée selon le type de base pour la règle [ϕ [x_1]] 132
Tableau 4.7 — Pourcentage des fonctions sémantiques des bases-sujets pour la règle [ϕ [x_1]] 134
Tableau 4.8 — Fonction sémantique des bases pour la règle [ϕ [x_1]] 135
Tableau 4.9 — Pourcentage de catégories grammaticales par argument pour la règle [ϕ [x_1]] 136
Tableau 4.10 — Pourcentage d'arguments par fonction syntaxique pour la règle [ϕ [x_1]] 139
Tableau 4.11 — Pourcentage d'arguments par fonction sémantique pour la règle [ϕ [x_1]] 141
Tableau 4.12 — Fonctions sémantiques de l'argument *somática* 143
Tableau 4.13 — Fonctions sémantiques de l'argument *floral* 144
Tableau 4.14 — Pourcentage de SA ou de SP par formule syntaxique pour la règle [ϕ [x_1]] 145
Tableau 4.15 — Changement de structures de l'argument pour un même concept 150
Tableau 4.16 — Relations conceptuelles et fonctions sémantiques pour la règle [ϕ [x_1]] 153
Tableau 4.17 — Fonctions sémantiques des satellites de la règle [[ϕ [x_1]] [y_1]] 165
Tableau 4.18 — Exemples d'UTC par formule morphologique pour la règle [[ϕ [x_1]] [y_1]] 167
Tableau 4.19 — Relations conceptuelles et fonctions sémantiques pour la règle [[ϕ [x_1]] [y_1]] 169
Tableau 4.20 — Fonctions sémantiques des satellites pour la règle [ϕ [y_1]] 172
Tableau 4.21 — Relations conceptuelles et fonctions sémantiques pour la règle [ϕ] [y_1]] 174
Tableau 4.22 — Exemples d'UTC formées d'arguments adverbiaux 180
Tableau 4.23 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1 [$x_{1/1}$]]] 184
Tableau 4.24 — Relations conceptuelles et fonctions sémantiques pour la règle[ϕ [x_1 [$x_{1/1}$]]] 186
Tableau 4.25 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 193
Tableau 4.26 — Relations conceptuelles et fonctions sémantiques pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] 195
Tableau 4.27 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1] [x_2]] 203
Tableau 4.28 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1] [x_2] [x_3]] 203
Tableau 4.29 — Relations conceptuelles pour la règle [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] 205
Tableau 4.30 — Relations conceptuelles et fonctions sémantiques de la règle [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]] 212
Tableau 4.31 — Relations conceptuelles pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]] 217
Tableau 4.32 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]] 221
Tableau 4.33 — Relations conceptuelles pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]] 221
Tableau 5.1 — UTC acceptées par le système 238

0. Introduction

Les termes représentent un élément essentiel pour l'expression des idées en terminologie. Dès la découverte d'un objet, une nouvelle notion surgit et par conséquent la création d'un nouveau terme est nécessaire pour exprimer la notion qui vient de naître. Toutefois, la relation établie entre le terme et la notion n'est pas toujours facile à comprendre.

En effet, quand le terme se combine avec d'autres pour former ce qu'on appelle une UTC, l'analyse s'avère plus complexe. L'analyse doit alors se faire non seulement en relation avec la notion de chaque composant, mais elle doit également prendre en considération la relation sémantique établie entre les composants. De cette façon, les composants d'une UTC sont perçus par rapport à leur fonction sémantique, ce qui permet de mieux définir leur emploi dans une UTC. C'est ainsi que le terme simple *planta* (plante), qui signifie « végétal multicellulaire » (v. Le Nouveau Petit Robert, 1995), lorsqu'il est employé dans l'UTC *clonagem de genes de plantas* (clonage de gènes de plantes), acquiert de nouvelles caractéristiques qui expriment des fonctions sémantiques qu'il n'avait pas à lui seul. Dans ce cas, *plantas* (plantes), vu dans une perspective fonctionnelle, a la fonction d'appartenance par rapport aux *genes* (gènes) qui sont, pour leur part, le résultat du processus de *clonagem* (clonage).

Nous voyons ainsi que c'est le rapport entre les composants qui est mis en relief dans ce type d'interprétation. Cette façon de procéder est fondée sur la manière propre aux spécialistes de nommer les « choses » en fonction de leur emploi dans un domaine

déterminé, ce qui rend son analyse assez complexe. C'est ainsi que la terminologie est qualifiée d'onomasiologique, puisqu'il s'agit d'une science qui étudie les relations qui vont de la « chose » aux signes linguistiques correspondants, c'est-à-dire une science qui étudie la sémantique des dénominations de la réalité.

C'est dans cette perspective de la description des phénomènes linguistiques réels qui gravitent autour d'une terminologie donnée et de la reconnaissance de l'usage des termes que nous avons fondé notre recherche. Pour ce faire, il a fallu choisir une théorie qui offrait les moyens nécessaires pour une analyse de cette nature. Nous avons ainsi opté pour la théorie de la grammaire fonctionnelle décrite par Simon Dik qui permettait d'en adapter les principes de base pour effectuer une analyse cohérente de la terminologie d'un domaine. Ces principes se résument à la conception de la langue comme un instrument d'interaction sociale et à la reconnaissance du rôle primordial de la pragmatique dans l'analyse de la langue.

Dans la pratique, cette théorie nous a fourni les bases pour l'élaboration, à un niveau plus abstrait, des règles de formation adaptées à l'analyse des UTC, ce qui a contribué énormément à la systématisation de nos données. Ces bases sont le fruit d'une vision de la langue comme un objet d'étude dynamique et variable et d'une « conception de la langue comme un système dans lequel s'insèrent des éléments nucléaires et des éléments périphériques et donc qui conçoit les règles et les principes de la grammaire davantage comme des tendances que comme des règles absolues avec des conditions de type oui/non d'application »[1] (Neves, 1997 :146). C'est justement la description et l'analyse des points fondamentaux de la théorie employée dans notre recherche qui constituent le sujet du premier chapitre.

Étant donné que notre recherche vise non seulement la description linguistique mais également l'utilisation de cette description pour la conception d'un logiciel d'extraction de termes en contexte, l'état de la question, présenté au chapitre deux, a tenu compte de ces

deux volets. Ainsi, la première partie du chapitre deux est consacrée aux principales études de base fonctionnaliste sur la description des UTC. Dans la seconde partie, nous avons effectué un survol des logiciels d'automatisation de l'extraction terminologique en mettant plus précisément en relief ceux qui ont été conçus dans une perspective fonctionnaliste.

Le chapitre deux traite également de la problématique, de la justification de la recherche et des objectifs. Ces trois aspects expliquent en détail ce qui a été l'élément déclencheur de notre recherche, soit les besoins de la Banque de données terminologiques du Brésil (Brasilterm).

Quant au chapitre trois, il a été dédié aux questions concernant la constitution et l'informatisation du corpus. En ce qui concerne la partie sur la constitution du corpus, nous avons touché aux aspects liés au choix de la biotechnologie des végétaux comme domaine de recherche ainsi qu'aux considérations qui nous ont amenée à définir la typologie des textes utilisée pour le corpus de collecte de données. Pour ce qui est de l'informatisation du corpus, nous avons décrit les divers types de logiciels employés pour le recensement des UTC. Dans cette partie, il a été question également de la description de la base de données créée à l'aide du logiciel *Access* pour systématiser les données de notre recherche.

Les données sont analysées et interprétées dans le chapitre quatre. Ce chapitre est divisé par type de règle de formation des UTC. Ces règles, au nombre de dix, ont été définies selon la structure de la prédication adaptée de la grammaire fonctionnelle. Pour chacune de ces règles, nous avons analysé les aspects grammaticaux, morphologiques, syntaxiques, sémantiques et pragmatiques des UTC relevées.

Finalement, le chapitre cinq présente plusieurs stratégies pour l'automatisation de l'extraction des UTC. Ces stratégies sont fondées sur les résultats obtenus dans l'analyse et l'interprétation des données. Il s'agit de propositions qui privilégient les environnements

[1] Traduction personnelle du texte original rédigé en portugais.

qui ont recours à l'intelligence artificielle et qui sont orientés vers la modélisation de connaissances linguistiques.

À partir d'une conception d'automatisation fondée sur l'élaboration d'une base de connaissances, nous proposons une structure générale du traitement automatique pour l'extraction des UTC ainsi que des pistes d'origine morphosyntaxique et sémantique pour le dépistage des UTC. Pour conclure, nous présentons une application virtuelle de nos pistes faite à partir de deux nouveaux paragraphes tirés d'un de nos textes.

La première annexe est réservée aux 210 UTC contenues dans la base de données et à leurs descriptions. La seconde annexe contient la liste des étiquettes utilisées pour représenter les fonctions sémantiques employées au cours de l'analyse des données, tandis que le texte qui a servi à tester les pistes de notre logiciel figure à la troisième annexe.

1. Cadre théorique

1.1 Introduction

Nous présenterons, dans les pages qui suivent, les fondements de notre recherche, c'est-à-dire les principes de base de la théorie de la grammaire fonctionnelle de Dik. Pour ce faire, nous allons d'abord situer l'école fonctionnaliste, pour en arriver ensuite à expliquer les principaux concepts du modèle théorique choisi. Finalement, nous ferons ressortir les adaptations que nous avons apportées au modèle pour qu'il réponde aux besoins de notre analyse de données et illustrerons le tout à l'aide d'exemples tirés du domaine de la biotechnologie des végétaux.

1.2 Linguistique fonctionnelle

Dans la littérature portant sur les origines de la linguistique fonctionnelle, on trouve souvent l'idée innovatrice de la linguistique saussurienne concernant le rôle de la langue comme instrument de communication. Toutefois, même si on ne peut pas considérer Saussure comme un fonctionnaliste, certains représentants de l'école fonctionnaliste ont repris son principe et « [...] considèrent l'étude d'une langue comme la recherche des fonctions jouées par les éléments, les classes et les mécanismes qui interviennent en elle » (Ducrot et Todorov, 1972 : 42). Il y a également des auteurs qui vont plus loin et qui voient la linguistique fonctionnelle comme une branche du structuralisme. C'est sur cette base que,

par exemple, Daneš caractérise l'école de Prague. Selon cet auteur, l'approche linguistique de l'école de Prague est principalement perçue comme un structuralisme fonctionnel. C'est la caractéristique dite fonctionnelle de l'analyse de la langue qui est souvent utilisée pour faire la distinction entre le structuralisme de Prague et les autres courants structuralistes (v. Daneš, 1987 : 3).

Dans ce contexte d'essai de démarcation entre les diverses écoles de la linguistique, Dirven et Fried expliquent ce qui différencie les courants linguistiques qui existent dans le monde. Ils distinguent ainsi l'axe fonctionnaliste et l'axe formaliste en disant que du côté fonctionnel « [...] the function of language forms seems to play the predominant role [...] » tandis que pour ce qui est du formalisme « [...] the analysis of linguistic form seems primary, whereas functional concerns are only secondary » (1987 : x).

Sur le plan historique, la linguistique fonctionnelle est née de l'école linguistique de Prague. Cette école a été créée en 1926 et a beaucoup influencé la linguistique européenne dans les années qui ont suivi la Seconde Guerre mondiale (v. Lyons, 1981 : 224). Dès le début, la notion de fonction a été la base théorique et méthodologique de la pensée de l'école de Prague. Cependant, cette notion a pris diverses acceptions au fur et à mesure que la théorie a été reprise par d'autres courants. Ainsi, même s'il existe aujourd'hui plusieurs approches de la conception fonctionnaliste de la langue, nous pouvons dire qu'elles ont en commun la vérification, la description et l'analyse des multiples façons de procéder de la langue pour atteindre l'objectif qui est de communiquer.

Autrement dit, ces approches ont été conçues différemment en gardant toutefois le même rapport entre la langue et la communication. De plus, comme le précisent Dirven et Fried, « [...] although the concept of functionalism in linguistics is explicity linked (or re-linked) with the Prague School of Linguistics, it should also be realized that it has influenced various others 'functional' approaches in linguistics in East and West and that functionalism has also started a life of its own, quite independently of the Prague School [...] » (1987 : ix).

Martinet est un exemple de représentant de la pensée fonctionnaliste européenne. Défenseur incontestable du fonctionnalisme en linguistique, cet auteur définit la langue comme un instrument ayant une fonction essentielle : la communication (v. 1991 : 9). En effet, c'est le besoin de communiquer qui fait de la langue un instrument d'interaction sociale. À ce sujet, Martinet affirme que « [...] si toute langue se modifie au cours du temps, c'est essentiellement pour s'adapter de la façon la plus économique à la satisfaction des besoins de communication de la communauté qui la parle » (*ibid*.). Pour Martinet, le mot *fonction* a un sens fondamental qui est celui de rôle que joue la langue dans la communication de l'expérience humaine (v. 1994 : 13). Ce postulat résume bien l'objet d'étude de la linguistique pour tous ceux qui s'inscrivent dans le courant fonctionnaliste.

Dans une vision plus globale, Mahmoudian, caractérise la linguistique fonctionnelle selon deux principes fondamentaux :

a) la conception instrumentale du langage;
b) la subdivision et la hiérarchisation fonctionnelle (v. 1979 : 3).

En ce qui concerne le premier principe, Mahmoudian dit que « [...] les langues sont considérées comme des outils servant à assurer des fonctions. Dépendant de ces fonctions, la forme de l'outil porte leurs empreintes, obéit à leurs impératifs; les langues sont structurées de façon à être adaptées à leurs fonctions » (*ibid*.). Il complète cette affirmation en précisant que, pour les fonctionnalistes, la fonction première des langues est la communication.

En ce qui a trait au deuxième principe, Mahmoudian explique que « [...] les parties de ces structures complexes que sont les langues sont distinguées suivant des critères fonctionnels, c'est-à-dire selon le rôle et la place qu'elles tiennent dans la communication linguistique » (*ibid.*). Il ajoute que l'importance de ces parties varie selon leurs fonctions dans la communication et que celles-ci agissent comme de critères de hiérarchisation.

Nous allons maintenant nous pencher sur la méthode de travail adoptée par la linguistique fonctionnelle. Celle-ci postule que l'étude de la langue doit se faire par rapport à la réalité où cette langue existe. Les diverses descriptions et analyses des éléments lexicaux simples ou complexes ont pour but de fournir des données pour la description du fonctionnement de la langue à l'intérieur de sa réalité. De cette façon, l'étude fonctionnaliste privilégie la pragmatique où la sémantique et la syntaxe sont analysées. Nous partons de la pragmatique pour arriver à la syntaxe par le biais de la sémantique.

En fait, dans la pratique, le chercheur fonctionnaliste récupère un élément lexical par l'analyse de son usage, de son contexte. Ainsi, le fonctionnaliste va toujours chercher le référent, la fonction de l'élément lexical dans un contexte.

Signalons également que la pensée fonctionnaliste a influencé divers domaines de la linguistique. Nous remarquons les efforts menés pour l'élaboration d'une théorie de la grammaire des langues. À ce sujet, la littérature récente attire l'attention sur deux modèles importants en grammaire fonctionnelle, celui de Halliday (Davidse 1987 : 39 - 79) et celui de Dik (1978; 1980; 1983; 1987; 1989).

Halliday, un représentant de l'école de Londres, et Dik, un représentant du groupe de Hollande, font partie du fonctionnalisme dit modéré. Ces deux auteurs ont défini la grammaire fonctionnelle comme une théorie qui propose l'organisation grammaticale des langues dans une perspective d'interaction sociale.

Comme l'explique Castilho, la grammaire fonctionnelle de Halliday met en évidence les usagers et l'utilisation de la langue et privilégie le récepteur et l'émetteur, aussi bien que la variation linguistique à partir de l'analyse grammaticale de la langue (v. 1994 : 76).

Quant à la grammaire fonctionnelle de Dik, elle a été influencée en partie par les idées de l'école linguistique de Prague. D'après Gebruers, le caractère fonctionnel et dynamique de la langue est défendu tant par l'école de Prague que par la grammaire fonctionnelle de Dik (v. 1987 : 129). Nous retrouvons d'ailleurs ces deux caractéristiques dans la méthode de la prédication, que Dik propose, pour la description linguistique. Dans la section qui suit, nous allons présenter la théorie de la grammaire fonctionnelle de Dik en tenant principalement compte des parties qui s'appliquent à notre recherche.

1.3 Théorie de la grammaire fonctionnelle de Dik

La théorie de la grammaire fonctionnelle de Dik a été exposée pour la première fois dans son livre intitulé *Function Grammar* publié en 1978. Cet ouvrage a donné lieu à beaucoup d'autres recherches menées dans le domaine. À cet égard, Hoekstra et *al*. mentionnent que « Since its presentation in 1978, FG[2] has met with many positive reactions as will be clear from the numerous references in the literature as well as from the increasing number of linguists working within this framework » (1981 : 3).

Par la suite, les linguistes qui se sont penchés sur la théorie de Dik, soit en l'appliquant à d'autres langues que l'anglais, soit en révisant ses concepts de base, y ont également apporté des modifications. À la lumière de ces réactions, ce dernier a écrit une nouvelle version de sa grammaire fonctionnelle. Ainsi, en 1989, l'ouvrage *The Theory of Functional Grammar* a été publié afin de systématiser ces dernières contributions. Cette deuxième édition porte, selon son auteur, sur « [...] the question of how the linguistic expressions of natural languages of any type can be described and explained in a way that is typologically, pragmatically, and psychologically adequate» (Dik, 1989 : xiv). Autrement dit, c'est une théorie qui s'intéresse à l'organisation des expressions linguistiques dans les langues.

[2] FG = Functional grammar (Note personnelle).

L'adéquation typologique signifie, pour Dik, qu'une théorie doit être capable de produire des grammaires pour tous les types de langues, en rendant compte, d'une façon systématique, des similarités et des différences de chacune de ces langues.

En ce qui a trait à l'adéquation pragmatique, l'auteur mentionne que, dans une grammaire fonctionnelle, les expressions linguistiques ne doivent pas être analysées isolément, mais toujours prises comme un outil utilisé par un émetteur dont le but est de produire une réaction délibérée de la part du récepteur. Il signale également que ces expressions linguistiques sont émises dans un contexte, qu'elles sont entourées d'autres expressions et qu'elles suivent des paramètres essentiels en ce qui concerne la situation du discours.

Finalement, à propos de l'adéquation psychologique, Dik souligne la nécessité d'analyser les expressions linguistiques comme des produits émis par des êtres humains qui ont tous des propriétés psychologiques particulières. Selon lui, une grammaire doit décrire le plus précisément possible les modèles psychologiques de la compétence et du comportement linguistiques.

1.3.1 Bases de la théorie

Dik a conçu sa théorie à partir de plusieurs principes de base. Dans le cadre de la présente recherche, nous allons nous concentrer sur les principes qui ont une application pratique directe pour l'établissement de la méthodologie utilisée dans notre étude.

Premièrement, la grammaire fonctionnelle observe la langue selon une perspective fonctionnelle, c'est-à-dire que la langue est vue comme un instrument d'interaction sociale dont la fonction est la communication. Dans ce sens, la base à partir de laquelle la pensée fonctionnaliste analyse la langue diffère de la vision formelle qui place la dimension

fonctionnelle au second rang ou qui n'en tient tout simplement pas compte. Ainsi, selon le point de vue formaliste, la langue est un ensemble de phrases dont la principale fonction est l'expression de la pensée et non pas nécessairement la communication. De plus, dans une grammaire formelle, la syntaxe est appliquée indépendamment de la sémantique et de la pragmatique de la langue. Par contre, dans une perspective fonctionnelle, la grammaire est conçue à partir d'une base pragmatique à l'intérieur de laquelle la sémantique et la syntaxe sont analysées. Cet aspect est repris par Hoekstra *et al.* lorsque ces derniers traitent de la grammaire fonctionnelle. Ces auteurs constatent que « Within FG [...] syntax is not regarded as an autonomous system, but as a system based on semantics» (1981 : 3).

Deuxièmement, la grammaire fonctionnelle accorde à la pragmatique un rôle primordial dans l'analyse de la langue. Dans ce contexte, la langue est considérée comme une activité structurée et coopérative. D'un côté, elle est structurée puisque les expressions linguistiques (concrétisations de la langue) obéissent à des règles sémantiques, syntaxiques, morphologiques, phonologiques et pragmatiques. De l'autre, elle est coopérative puisqu'elle fait interagir au moins deux participants qui essaient de rendre la communication la plus efficace possible. En bref, les règles relatives à la construction des expressions linguistiques, soit les règles sémantiques, syntaxiques, morphologiques et phonologiques, sont considérées comme des outils pour rendre l'utilisation de la langue (la dimension pragmatique) possible. Ce paradigme fonctionnel est repris par Dik lorsqu'il formule le principe selon lequel « [...] a theory of language should not be content to display the rules and principles underlying the constructions of linguistic expressions for their own sake, but should try, wherever this is possible at all, to explain these rules and principles in terms of their functionality with respect to the way in which these expressions are used» (1989 : 4).

Troisièmement, dans un contexte d'ordre méthodologique, la grammaire fonctionnelle propose la construction de prédications abstraites à partir de structures déjà présentes dans le lexique, ce que Dik appelle la méthode de la prédication pour la description linguistique. C'est ainsi que dans un article sur la grammaire fonctionnelle,

Neves explique qu'à partir du modèle de Dik, tous les éléments lexicaux d'une langue doivent être analysés au moyen de la prédication (v. 1994 : 120). Ces prédications sont révélées par la construction de structures qui spécifient les propriétés fonctionnelles et catégorielles de chaque constituant d'une expression linguistique. Toutefois, comme le fait remarquer Dik, même si les propriétés catégorielles et fonctionnelles sont capitales pour la description, la grammaire fonctionnelle privilégie les propriétés qui caractérisent les relations fonctionnelles entre les éléments de l'expression linguistique analysée (v. 1978 : 13). Dik ajoute que

> Linguistic expressions are complex networks, characterized by functional relations operative at different levels. Functions are also needed (alongside categories) because functions and categories do not stand in a one-to-one relation to each other. The same category may occur in different functions, and the same function may apply to constituents with different categorial properties (1989 : 24).

Afin de clarifier les concepts propres à la méthode de la prédication examinée brièvement ci-dessus, nous allons les décrire en détail dans les pages qui suivent. Comme nous l'avons déjà mentionné, notre principale intention a été de présenter la méthode proposée par Dik et de l'adapter à notre recherche.

1.3.2 Méthode de la prédication

La méthode de la prédication présentée par Dik (1978; 1981; 1983; 1989) permet d'analyser les expressions linguistiques qui existent dans les langues. Elle utilise le modèle de prédication composée de prédicats et de termes. Les prédicats sont des expressions qui désignent des propriétés (p. ex. les catégories grammaticales) ou des relations entre les entités (p. ex. les fonctions sémantiques) (v. Dik, 1983 : 10). La prédication est donc l'application de quelques prédicats à un nombre donné de termes. Ces termes peuvent être soit des bases, soit des termes dérivés (v. Dik, 1981 : 4 et 6). L'ensemble des *prédicats de base* est composé de toutes les structures prédicatives d'une langue, tandis que les *prédicats*

dérivés sont l'application de ces structures au lexique d'une langue. Afin de bien comprendre ce que la grammaire fonctionnelle entend par prédicat, Dik adopte quelques considérations de base. Parmi celles-ci, nous avons retenu uniquement celles qui nous semblent les plus pertinentes pour notre recherche :

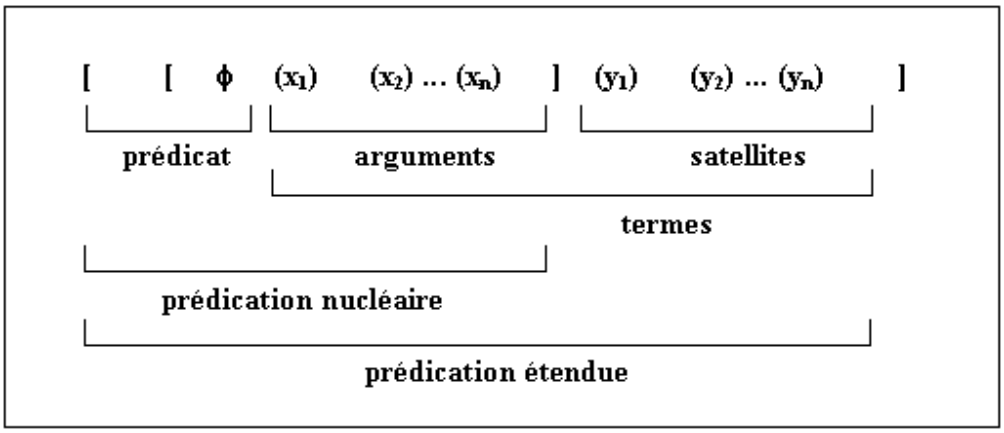
a) tous les éléments lexicaux d'une langue sont analysables comme prédicats;
b) il peut exister différents types de prédicats (verbal, adjectival ou nominal);
c) tous les prédicats sont sémantiquement interprétés comme désignant des propriétés ou des relations;
d) les prédicats sont analysés par des structures de prédicats qui spécifient leurs propriétés sémantiques et syntaxiques (v. 1989 : 54).

Ainsi, la grammaire fonctionnelle présente un modèle théorique qui permet d'établir la structure fonctionnelle des expressions linguistiques par le biais d'un prédicat inséré dans des prédications.

1.3.2.1 Structure de la prédication

Une structure de prédicat devient une prédication quand des termes sont appliqués à cette structure (v. Dik,1981 : 7). Voici comment Dik représente la structure générale de la prédication (v. 1978 : 26).

Figure 1.1 — Structure générale de la prédication



Examinons chacune des étiquettes du modèle de Dik :

Prédicat : Le prédicat est le centre de l'expression linguistique. C'est à partir de cet élément que nous pouvons décrire les prédications nucléaire et étendue. Le prédicat comporte toute l'information nécessaire pour la description de la prédication.

Arguments : Les arguments sont des éléments exigés par la sémantique de la prédication. Ils sont fondamentaux pour la constitution de la prédication nucléaire. Entre autres, les arguments ont une relation directe avec le prédicat et ils occupent une position essentielle dans la prédication.

Satellites : Les satellites sont des éléments qui n'entrent pas dans la définition des états de chose[3] comme tel, mais qui donnent plus d'information sur ces états de chose. Par exemple, les satellites peuvent informer sur la localisation d'un état de chose. Par opposition aux arguments, les satellites ont une relation avec la prédication nucléaire prise comme un tout et ils occupent, de

[3] Selon Dik, l'état de chose « is here used in the wide sense of "conception of something which can be in some world" » (1989 : 89).

préférence, une position périphérique dans l'expression linguistique.

Termes : Les termes sont des expressions qui font référence aux entités d'une réalité. Ils occupent les positions d'arguments ou de satellites dans la structure de la prédication.

Prédication nucléaire : La prédication nucléaire est l'application d'un prédicat à un nombre donné de termes qui fonctionnent comme arguments de ce prédicat. La prédication nucléaire est formée d'un prédicat et de ses relations avec un nombre déterminé de termes qui ont la fonction d'arguments. La prédication nucléaire définit un ensemble d'états de chose.

Prédication étendue : La prédication étendue est constituée de la prédication nucléaire et de satellites.

1.3.2.2 Différences entre arguments et satellites

Comme nous l'avons mentionné précédemment, ce qui distingue les arguments et les satellites c'est l'élément avec lequel ils ont un rapport. C'est-à-dire que, si, d'une part, les arguments ont une relation avec les prédicats, d'autre part, les satellites ont un rapport avec la prédication nucléaire prise comme un tout.

Par conséquent, leurs positions dans l'expression linguistique varient conformément à cette caractéristique relationnelle. Ainsi, dans la pratique, les arguments occupent une position plus centrale à l'intérieur de la prédication que les satellites.

Toutefois, bien que la distinction entre les arguments et les satellites semble simple à première vue, il n'en est rien quand nous essayons de l'appliquer à une réalité

linguistique. En effet, il faut noter que la ligne de démarcation entre ces deux éléments n'est pas toujours très nette. Cette observation est confirmée par Dik qui affirme que « [...] in certain cases it may be difficult to decide whether a given term has the status of an argument or a satellite » (1989 : 73).

Considérons le cas de *in Amsterdam* dans les exemples suivants extraits de Dik (v. 1989 : 73) :

a. *John bought a car* ***in Amsterdam***
b. *John lives* ***in Amsterdam***

Dans les deux exemples *in Amsterdam* a une fonction sémantique de localisation, mais cet élément est un satellite dans le premier cas, tandis que dans le second il a un statut d'argument. Plus précisement, dans (a) *in Amsterdam* a un rapport avec tout ce qui vient avant (*John bought a car*) alors que dans (b) il s'agit essentiellement d'un argument du prédicat *lives*.

Dik propose un moyen pour bien faire la distinction entre les arguments et les satellites. Selon lui, « [...] a satellite can be left out without affecting the grammaticality or the meaning of the remaining construction, whereas leaving out an argument will either render the remainder ungrammatical or change its semantics » (1989 : 74). Il n'en demeure pas moins que ce moyen ne fonctionne pas toujours.

C'est dans ces conditions que Dik analyse les satellites comme étant des éléments optionnels. Il mentionne que « When the satellite is not specified, we do not feel that something is missing which should be reconstructed on the basis of context and situation » (*id.* : 193).

Quelle que soit la distinction faite entre les arguments et les satellites, l'importance de ces éléments demeure liée à l'interprétation de la relation établie entre eux et le prédicat

dans l'expression linguistique. Dans la grammaire fonctionnelle de Dik, cette interprétation est obtenue par l'analyse des fonctions syntaxiques, sémantiques et pragmatiques exercées par des prédicats, des arguments et des satellites. Il sera d'ailleurs question de ces fonctions dans le chapitre qui suit.

1.3.2.3 Fonctions de la prédication

Les informations contenues dans chaque structure du prédicat sont de nature multiple. Au-delà des informations sur l'existence d'un prédicat, des arguments, des satellites et de leur localisation à l'intérieur de l'expression linguistique, la structure du prédicat renseigne également sur les diverses fonctions syntaxiques, sémantiques et pragmatiques établies entre le prédicat et les arguments ou entre la prédication nucléaire et les satellites.

Brito reprend les fonctions de Dik et les décrit comme suit :

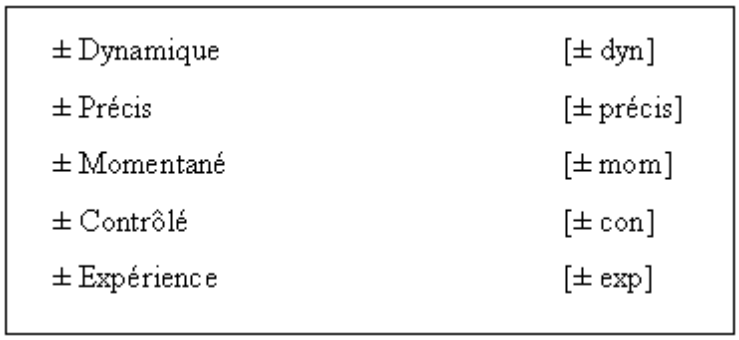
a) les fonctions sémantiques spécifient les rôles que chaque référent, représenté par les termes, a par rapport aux états de chose : *agent*, *but*, *contenant*, pour n'en nommer que quelques-uns;
b) les fonctions syntaxiques spécifient la perspective selon laquelle l'état de chose est présenté dans l'expression linguistique : *sujet* et *objet*;
c) les fonctions pragmatiques spécifient le statut informationnel de chaque élément d'une prédication : *thème, topique, focus*, etc. (v. 1994 : 82).

L'origine de ces fonctions varient : les fonctions sémantiques dépendent du rôle exercé par le lexique lui-même, tandis que les fonctions syntaxiques et pragmatiques dérivent du contexte où figure l'expression linguistique.

C'est donc le rôle du lexique qui détermine les informations sémantiques sur l'état de chose de l'expression linguistique analysée. Étant donné le rapport étroit entre la sémantique et la syntaxe dans une étude de nature fonctionnaliste, il est évident que la syntaxe contribue directement à la délimitation des rôles sémantiques joués par le lexique.

Dik fixe des paramètres pour établir une typologie des fonctions sémantiques (v. 1989 : 89 - 109). Ces paramètres sont résumés dans la figure qui suit.

Figure 1.2 — Paramètres pour l'établissement des fonctions sémantiques



Dans le cas de d*ynamique*, il y a un rapport avec la notion de changement qui peut être positif ou négatif. En ce qui concerne le [- dyn] il n'y a pas de changement, ce que Dik appelle s*ituation*. Inversement, en ce qui a trait au [+ dyn] nous remarquons un dynamisme et Dik le nomme é*vénement*. Considérons les exemples suivants :

a. *John **was sitting** in his father's chair* [- dyn]
b. *John **opened** the door* [+ dyn] (*id.* : 91).

Dans le premier exemple, nous pouvons dire que le rapport est moins dynamique [- dyn] parce qu'il n'y a aucun changement. L'agent, *John*, est demeuré assis sur la chaise de son père. Dans le second exemple, l'agent effectue une action (ouvrir la porte) qui implique l'idée de mouvement, donc un changement.

Pour ce qui est du paramètre *précis*, il indique avec plus ou moins de précision que l'é*vénement* va se produire. Dik donne les exemples suivants pour expliquer ce paramètre :

a. *John walked* ***in the park*** [- précis]
b. *John walked* ***to the station*** [+ précis] (*id.* : 92).

À notre avis, dans ces exemples, ce sont les prépositions qui donnent toute l'information quant à la précision de l'*événement.* En effet, ce n'est ni *park* ni *station* qui fournissent le degré de précision de l'*événement*. C'est plutôt la composante prépositionnelle liée au verbe qui a toute la charge sémantique quant au degré de précision. Ainsi, *walked to* nous mène à un endroit exact (dans ce cas-ci la *station*), tandis que *walked in* désigne un endroit indéfini du *park*. Si on inverse les prépositions, le résultat sera le suivant.

a. *John walked* ***to the park*** [+ précis]
b. *John walked* ***in the station*** [- précis].

Le paramètre *momentané* s'applique aux *événements* de type ponctuel, tandis que [- mom] est réservé aux *événements* qui ont une certaine durée. Pour distinguer ces deux paramètres, Dik précise que les é*vénements* [- mom] ne peuvent être combinés qu'avec des verbes « [...] which signal the begining, the continuation, or the end of the Event » (*id.* : 95). Ainsi les exemples suivants présentés par Dik :

a. *John* ***started / continued / finished*** *painting the portait* [- mom]
b. **John* ***started / continued / finished*** *reaching the summit* [- mom] (*ibid.*).

Le paramètre *contrôlé* renvoi à la notion de ce qui contrôle la *situation* ou l'*événement*. Dans le premier exemple cité ci-dessous, nous avons un *événement* contrôlé [+ con], puisque l'*agent* a le plein contrôle de l'action qu'il pose, tandis que le deuxième est

caractérisé par une *situation* non contrôlée [- con] parce que l'entité n'a aucun contrôle sur le *processus*. Voici les deux exemples pris de Dik :

a. ***John*** *opened the door* [+ con]
b. ***The substance*** *was red* [- con] (*id.* : 96).

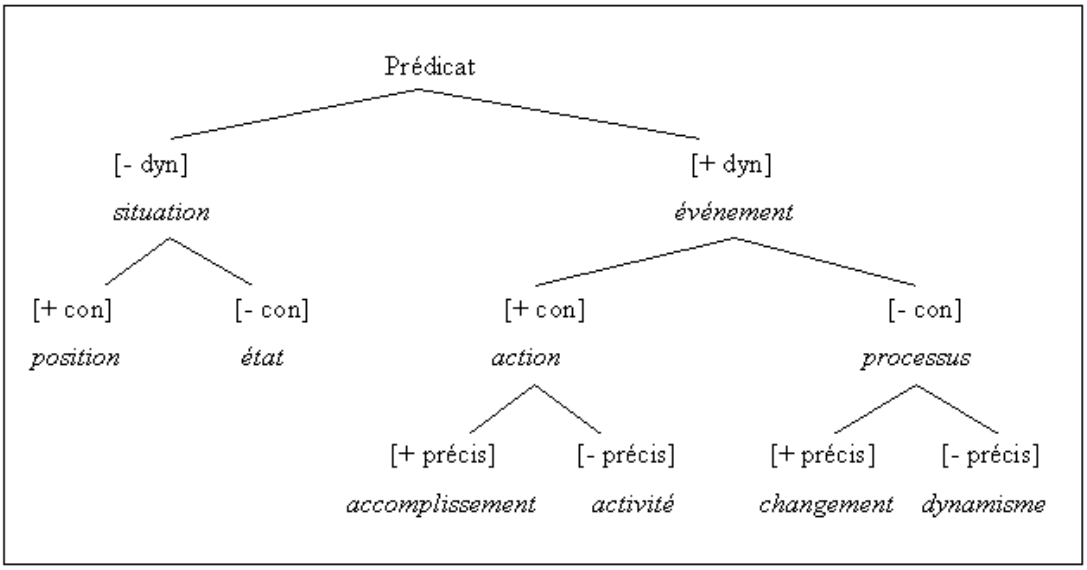
Finalement, le dernier paramètre appelé e*xpérience* fait référence aux prédicats qui reflètent des perceptions, des sentiments, des volontés, des conceptions ou, autrement dit, des expériences humaines. Voici deux exemples qui illustrent ce paramètre :

a. *John* ***got an interesting idea*** [+ exp]
b. *John* ***dreamed about his girl friend*** [+ exp] (*id.* : 99).

Toutefois, Dik ne tient pas vraiment compte de ce paramètre et conclut même que « [...] the oppostion [± exp] does not necessarily have a deep impact on the grammatical organization of natural languages » (*ibid.*).

Afin de bien comprendre toutes ces notions, nous avons reproduit ci-dessous le schéma présenté par Dik pour illustrer les interrelations entre les divers paramètres (*id.* : 98).

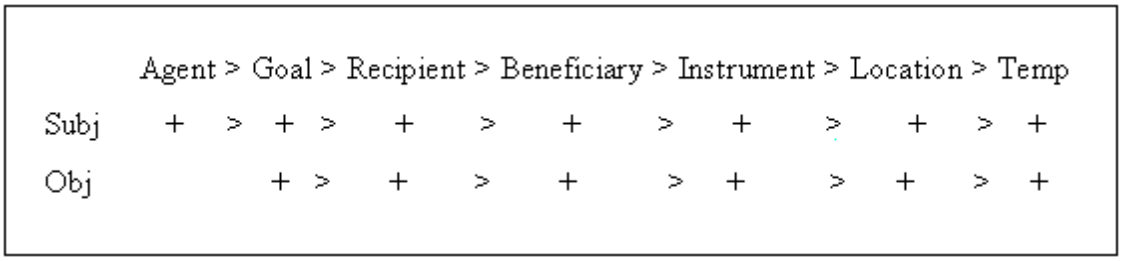
Figure 1.3 — Schéma des interrelations entre les paramètres



À partir de ce schéma, nous voyons que, par exemple, une *action* est un *événement* plus contrôlé tandis qu'un *processus* se caractérise comme étant un *événement* moins contrôlé. Ce type de distinction est très utile lorsque vient le temps d'attribuer des fonctions sémantiques aux divers éléments d'une expression linguistique.

Étant donné que la grammaire fonctionnelle de Dik se fonde sur une théorie qui envisage la syntaxe par le biais de la sémantique, l'interprétation des fonctions syntaxiques est évidemment intégrée aux fonctions sémantiques. C'est ainsi que Dik lie l'attribution des fonctions syntaxiques dans l'expression linguistique à un type de hiérarchie. Cette hiérarchie est, en réalité, de nature sémantique. Dik la nomme *hiérarchie de la fonction sémantique* (v. *id.* : 223). En fait, la grammaire fonctionnelle part du principe que chaque élément d'une prédication a une fonction sémantique et qu'il occupe, selon son importance, des positions plus ou moins centrales dans la prédication. Ce degré d'importance, exprimé par Dik sous forme de hiérarchie (v. *id.* : 226), est reproduit dans la figure ci-dessous.

Figure 1.4 — Hiérarchie de la fonction sémantique



Selon Dik, « the idea is that as we proceed through the SFH[4] from the more "central" to the more "peripheral" semantic functions, Subj and Obj assignment become more and more "difficult", and the resulting constructions become more and more "marked" » (*ibid.*).

Il faut signaler que pour Dik les fonctions syntaxiques *sujet* et *objet* ont un autre sens que celui de la grammaire traditionnelle[5]. D'après lui,

> [...] the notions Subject and Object as used in FG will undergo a reinterpretation in such a way they will be regarded as having their own contribution to the semantics of the expression, a contribution consisting in defining different "perspectives" over the state of affairs designated by the predication. For that reason, "perspectival functions" might be a better term to cover their essential nature (*id.* : 25).

Effectivement, dans la grammaire fonctionnelle, la fonction syntaxique *sujet*, par exemple, est définie selon une base sémantique, autrement dit la fonction *sujet* est attribuée aux éléments qui font référence à des entités « taken as a point of departure for the presentation of the state of affairs in which it participates » (Dik 1978 : 87).

À propos de la fonction pragmatique, nous remarquons que la grammaire fonctionnelle de Dik lui réserve un rôle primordial. De toute évidence, cette attitude reflète

[4] SFH signifie *semantic function hierarchy*.

[5] La grammaire traditionnelle définit le *sujet* comme étant celui qui fait ou subit l'action exprimée par le verbe et le *complément d'objet* comme étant le syntagme nominal complément du verbe qui subit l'action faite par le sujet (v. Dubois, 1991 : 344 et 467).

le principe fonctionnel selon lequel la pragmatique sert de fondement pour l'analyse de la sémantique et de la syntaxe.

Dans la perspective de la grammaire fonctionnelle, ce sont les fonctions pragmatiques qui fournissent

> [...] the informational status of the constituents, in relation to the wider communicative setting in which they are used. The main parameters along with [...] pragmatic functions can be distinguished are "topicality" (= characterizing "the things we talk about") and "focality" (= characterizing the most important or salient parts of what we say about the topical or things) (Dik 1989 : 264).

Ainsi, la fonction pragmatique *topique* est attribuée aux éléments dans l'expression linguistique à laquelle on fait référence, c'est-à-dire à la partie centrale sur laquelle on dit quelque chose.

En ce qui a trait à la fonction *focus*, elle est assignée aux éléments qui présentent ce qui est plus important ou plus prédominant sur le *topique*.

Tout compte fait, c'est par le biais de l'application de ces fonctions (sémantiques, syntaxiques et pragmatiques) que nous parvenons à formuler des structures de prédicats.

1.4 Application de la théorie de la grammaire fonctionnelle de Dik

La théorie de la grammaire fonctionnelle de Dik a été conçue pour l'analyse de la langue générale, plus précisement pour étudier la structure interne des phrases ainsi que leurs fonctions sémantiques, syntaxiques et pragmatiques. À notre connaissance, cette théorie n'a jamais encore été utilisée à des fins d'analyse en langue de spécialité.

Toutefois, les auteurs qui se sont penchés sur la structure interne des mots, l'interprétation sémantique et pragmatique ainsi que le traitement automatique des unités terminologiques complexes (dorénavant UTC), ont mis à profit certains principes généraux de l'école fonctionnaliste. Nous reviendrons sur ce sujet au chapitre suivant.

Nous allons maintenant exposer les adaptations que nous avons faites à la théorie de Dik pour pouvoir l'appliquer lors de l'analyse de notre corpus qui traite de la biotechnologie des végétaux.

1.4.1 Adaptations de base

Depuis les années 30, la terminologie a vécu un changement de perspective. D'une discipline prescriptive et normalisatrice, elle est devenue un phénomène envisagé de plus en plus selon un point de vue descriptif et fonctionnel.

Dans ce nouveau contexte, le vocabulaire spécialisé est perçu comme un objet d'analyse fonctionnelle. Par conséquent, la langue de spécialité est actuellement analysée dans son contexte, autrement dit dans la situation pragmatique du discours. Cette attitude a des conséquences multiples sur la façon d'étudier la langue, comme l'acceptation de l'existence de variations dans les lexiques. À ce sujet, Faulstich fait remarquer que « [...] la dimension pragmatique discursive révèle le niveau de variation des termes dans le corpus spécialisé et, en même temps, révèle le fonctionnement de la terminologie dans le milieu communicatif où elle a été créée » [6] (1997 : 17).

Pour ce qui est de la conception de la langue, le terminologue fonctionnaliste envisage la terminologie dans des perspectives analogues à la grammaire fonctionnelle qui définit la langue générale comme un instrument d'interaction sociale dont le but est de communiquer.

[6] Traduction personnelle du texte original rédigé en portugais.

De plus, en ce qui concerne le traitement de la sémantique, de la syntaxe et de la pragmatique en terminologie, les travaux fonctionnalistes suivent le même principe que la grammaire fonctionnelle, c'est-à-dire qu'ils mettent à profit l'étude de la syntaxe fondée sur la sémantique et analysée selon une base pragmatique.

Pour la présente recherche, nous avons adapté quatre principes de base de la grammaire fonctionnelle de Dik, sans toutefois nous éloigner de ses principaux fondements :

a) la règle de formation;
b) les fonctions sémantiques;
c) les fonctions syntaxiques;
d) les fonctions pragmatiques.

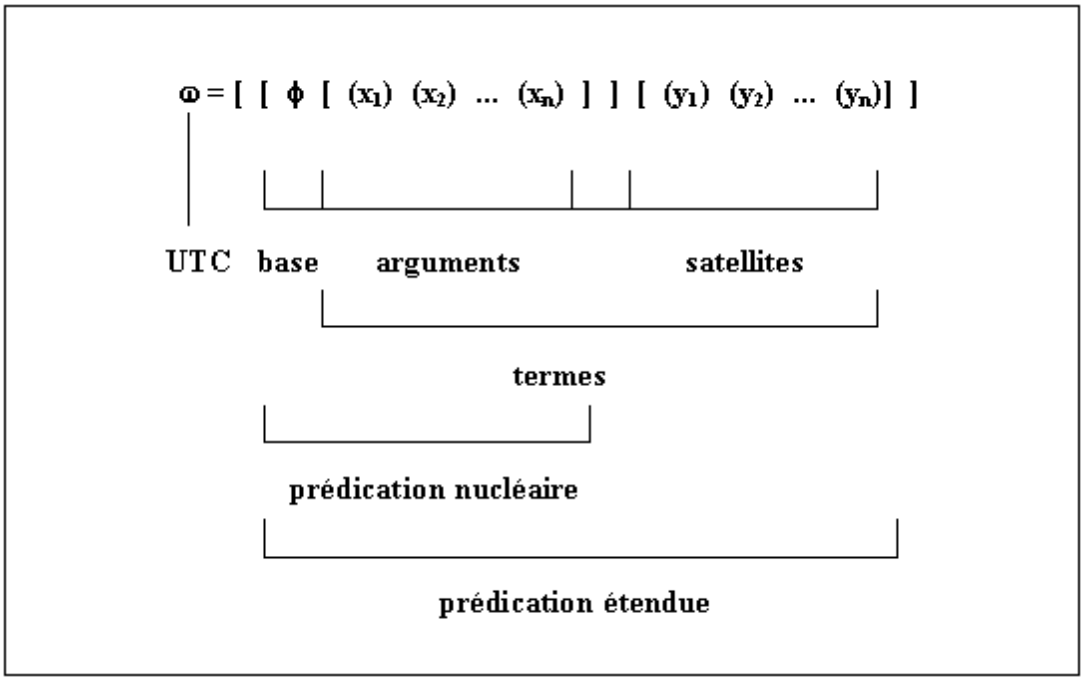
1.4.1.1 Règle de formation

Avant de définir la règle de formation adoptée pour notre recherche, il convient de préciser que notre but n'est pas l'analyse des phrases, mais des éléments lexicaux complexes qui figurent dans notre corpus.

Comme nous l'avons mentionné auparavant, la langue générale, pour la grammaire fonctionnelle, est formée de prédicats et de termes (arguments et satellites). Du côté des langues de spécialité, les UTC sont définies comme des unités formées d'une base et d'arguments ou d'une prédication nucléaire et de satellites. Ce sont ces deux groupes d'éléments qui constituent la structure générale de la prédication soit en langue générale (prédicats + arguments ou prédication nucléaire + satellites), soit en langue de spécialité

(base + arguments ou prédication nucléaire + satellites). Ainsi, suivant l'approche de la grammaire fonctionnelle, nous avons formulé une règle de formation des UTC (v. figure ci-dessous) adaptée à notre recherche.

Figure 1.5 — Règle de formation des UTC



où :

ω = UTC

ϕ = base de l'UTC

x_1, x_2, x_n = arguments

y_1, y_2, y_n = satellites

termes = arguments, satellites ou les deux à la fois

prédication nucléaire = base + arguments

prédication étendue = prédication nucléaire + satellites

Examinons un cas concret afin de bien comprendre la règle de formation présentée. Prenons par exemple l'UTC *aclimatação de plantas no solo* (acclimatation de plantes au sol). Cette UTC sera représentée de la façon suivante :

$$\omega = [\ [\ \phi\ [\ x_1\ ]\ ]\ [\ y_1\ ]\ ]$$

où:

ω (UTC) = *aclimatação de plantas no solo* (acclimatation de plantes au sol)
ϕ (base de l'UTC) = *aclimatação* (acclimatation)
x_1 (argument) = *de plantas* (de plantes)
y_1 (satellite) = *no solo* (au sol)

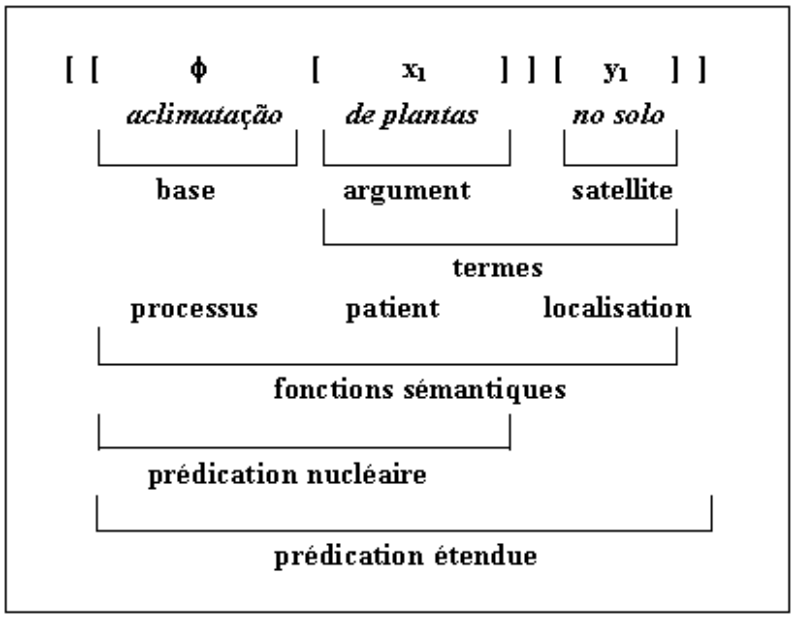
Une UTC a nécessairement une prédication nucléaire formée d'une base et d'au moins un argument. Les arguments représentés par la variable *x* fournissent aussi de l'information sur leur position dans l'UTC grâce à l'indice qui va de *1* (premier argument) jusqu'à *n*. De plus, l'UTC peut avoir des satellites qui occuperont une place dans la prédication étendue s'il y a lieu. La prédication étendue est représentée par la variable *y* et contient également un indice qui précise sa position à l'intérieur de l'UTC.

La base d'une UTC peut être considérée comme le noyau de l'unité. C'est à partir de la base que nous pouvons déduire les informations sémantiques, syntaxiques et pragmatiques des autres éléments de l'UTC, comme les arguments et les satellites. Tel est le cas de la base *aclimatação* (acclimatation) dans l'exemple précédent. En effet, *de plantas* (de plantes) a une fonction sémantique du type *patient* parce que nous savons dès le départ que la fonction sémantique de la base *aclimatação* (acclimatation) est un *processus*. Ainsi, *aclimatação* (acclimatation) est un *processus* qui agit sur l'argument *de plantas* (de plantes) qui est pour sa part un *patient* de ce *processus.* Le même mécanisme d'interprétation s'applique dans le cas des fonctions syntaxiques et pragmatiques.

Toutefois, il faut mentionner que les arguments et les satellites agissent également sur la base en lui apportant des précisions. Cette incidence entre les arguments et la base et entre les satellites et la prédication nucléaire est également représentée dans la règle de formation.

Comme nous pouvons le constater dans l'exemple présenté auparavant, les éléments *de plantas* (de plantes) et *no solo* (au sol) ont pour fonction de restreindre le signifié de l'UTC. Cette fonction de restriction varie selon qu'il s'agit d'arguments ou de satellites. D'une part, l'argument *de plantas* (de plantes) agit directement sur la base formant ainsi la prédication nucléaire *aclimatação de plantas* (acclimatation de plantes). D'autre part, le satellite *no solo* (au sol), qui a la fonction sémantique *localisation*, influe sur la prédication nucléaire au complet, formant la prédication étendue *aclimatação de plantas no solo* (acclimatation de plantes au sol). D'ailleurs, l'une des principales caractéristiques des satellites, est justement d'avoir un rapport avec la prédication nucléaire et non seulement avec la base comme c'est le cas pour un argument. Afin de résumer tous ces concepts, nous les avons présentés dans la figure qui suit.

Figure 1.6 — Représentation de la règle et des fonctions sémantiques de l'UTC



Nous avons décidé d'employer la règle de formation que nous venons de présenter pour systématiser les données de notre recherche. Au fur et à mesure que nous l'avons appliquée, nous avons obtenu plusieurs éclaircissements sur le phénomène de la prédication syntactico-sémantique des UTC.

1.4.1.2 Fonctions sémantiques, syntaxiques et pragmatiques

Dans la perspective de la grammaire fonctionnelle, l'interprétation sémantique, syntaxique et pragmatique de la base et de chaque argument ou satellite nous fournit les propriétés et les relations à l'intérieur d'une UTC. Ainsi, nous pouvons avoir différentes fonctions sémantiques, syntaxiques et pragmatiques selon les propriétés fonctionnelles de chaque élément.

Castilho analyse les rapports entre la prédication et la sémantique et affirme que « la prédication est un processus qui produit des signifiés qui n'existent pas à l'intérieur du sens

des éléments lexicaux, et qui dépend de la relation entre un élément-prédicatif et un élément-sujet »[7] (1994 : 81).

Au sujet de la formation des UTC, Faulstich ajoute que la prédication syntactico-sémantique « [...] correspond à la relation entre l'élément-prédicatif, représenté par un adjectif, par une locution qui commence par une préposition ou un adverbe, et un substantif qui est le noyau sémantique de l'UTC appelé sujet »[8] (1995 : 4).

De cette façon, et dans la perspective syntactico-sémantique, l'élément-prédicatif confère à la base de l'UTC une propriété qui, selon Castilho, peut être :

a) un changement de l'extension du signifié de la classe-sujet;
b) un changement des propriétés intensionnelles de la classe-sujet (v. 1994 : 81).

Si nous prenons l'UTC *mecanismo biológico de resistência* (mécanisme biologique de résistance), nous constatons que :

a) le nom *mecanismo* représente la base de l'UTC analysée;
b) cette base est suivie de l'adjectif *biológico* qui constitue le premier argument de l'UTC. Cet argument restreint le signifié de *mecanismo*, formant ainsi une première UTC : *mecanismo biológico*;
c) une nouvelle expansion à droite, par l'ajout d'un deuxième argument représenté par le syntagme prépositionnel *de resistência*, vient compléter l'UTC *mecanismo biológico de resistência.*

Ces expansions successives créent de nouvelles dimensions sémantiques et des notions plus spécifiques à l'UTC. La base *mecanismo* a reçu de nouvelles propriétés qui

[7] Traduction personnelle du texte original écrit en portugais.
[8] Traduction personnelle du texte original écrit en portugais.

n'existaient pas dans sa forme simple, c'est-à-dire que la base *mecanismo* a acquis d'autres propriétés intensionnelles et extensionnelles.

Mentionnons également que, dans l'espace syntagmatique de l'UTC, l'adjectif *biológico* a pris le nom *mecanismo* comme sujet. En outre, le syntagme prépositionnel *de resistência*, qui a pour fonction de spécifier la notion de *mecanismo biológico*, a pris comme sujet la base de l'UTC (*mecanismo*) plus son premier argument (*biológico*).

Ces considérations nous ont incitée à réfléchir sur l'importance de la définition des fonctions sémantiques, syntaxiques et pragmatiques spécifiques pour mieux comprendre le comportement linguistique des UTC. Ainsi, toujours suivant la théorie de la grammaire fonctionnelle, nous avons attribué des fonctions sémantiques, syntaxiques et pragmatiques pour chaque élément qui compose l'UTC. Ces fonctions ont été déterminées au fur et à mesure de l'analyse des UTC. Nous avons dû faire un ajustement à cette théorie parce que, comme nous l'avons déjà précisé, nous n'étudions pas des phrases, mais des unités complexes qui appartiennent à la langue de spécialité.

Nous avons donc mis à jour les fonctions sémantiques conformément à nos besoins. Voici la liste des fonctions sémantiques utilisées dans la présente recherche : *action, agent, appartenance, but, composant, contenance, effet, entité, état, forme, groupement, instrument, intensité, localisation, mode, patient, processus, propriété, résultat, source.*

Dans un premier temps, ces fonctions sont décrites pour chaque élément de l'UTC en tenant compte des rapports entre eux. Cette façon de faire nous aide à établir, dans un second temps, les différents types de relations conceptuelles qui existent entre la base et les arguments et entre la prédication nucléaire et les satellites. Le but final est d'arriver à établir une sorte de réseau conceptuel qui représenterait toutes les idées exprimées par les fonctions et les rapports sémantiques des UTC. Pour cette seconde étape, nous nous appuyons notamment sur Sowa qui définit un réseau sémantique comme « The collection of

all the relationships that concepts have to other concepts, to percepts, to procedures, and to motor mechanisms [...] » (1984 : 76).

Nous avons relevé quatre types de fonctions syntaxiques, à savoir : *sujet*, *complément du nom*, *adjoint de nom* et *complément circonstanciel*.

En ce qui concerne la fonction syntaxique *sujet*, nous nous éloignons, comme le fait la grammaire fonctionnelle, de la grammaire traditionnelle. À partir d'une analyse de base sémantique, nous réservons la fonction *sujet* aux éléments qui servent de point de départ à l'analyse d'une UTC. Autrement dit, le *sujet* est ce qui centralise la charge informationnelle du point du vue du comportement syntaxique de l'UTC prise comme un tout parce que déterminer le *sujet* à l'intérieur d'une UTC est une question de perspective d'analyse.

De plus, la fonction *sujet* ne doit pas être considérée comme nécessairement liée à une seule fonction sémantique. Dans ce contexte, Dik remarque que

> The interpretation of Subject may also be aided by the fact that FG does not identify it with the Agent on the semantic side, nor with Topic or any other pragmatic function on the pragmatic side, even though in actual linguistic expressions the functions of Agent and Subject, of Subject and Topic, or of Agent, Subject, and Topic will often coincide in the same constituent (1978 : 87).

Dans la terminologie de la biotechnologie des végétaux, nous avons relevé plusieurs UTC où la fonction *sujet* est représentée par un élément qui a une autre fonction sémantique que celle d'a*gent*. Le tableau 1.1 réunit quelques exemples qui illustrent la diversité des fonctions sémantiques pour la fonction syntaxique *sujet*.

Tableau 1.1 — Fonctions syntaxiques et sémantiques des UTC

UTC	Base	Fonction syntaxique	Fonction sémantique
agente redutor (agent réducteur)	*agente* (agent)	*sujet*	*agent*
agregado de células (agrégat de cellules)	*agregado* (agrégat)	*sujet*	*groupement*
árvore elite adulta (arbre élite adulte)	*árvore* (arbre)	*sujet*	*entité*
propagação clonal rápida (propagation clonale rapide)	*propagação* (propagation)	*sujet*	*processus*
sítio ativo (site actif)	*sítio* (site)	*sujet*	*localisation*

En ce qui concerne la fonction *complément du nom*, la *Nova gramática do português contemporâneo* la nomme « *complemento nominal* » et la définit comme un élément lié par une préposition à un substantif, un adjectif ou un adverbe qui en intègre ou en limite le sens (v. Cunha et Cintra, 1985 : 135). Cette définition correspond, *grosso modo*, à celle que Grevisse donne à l'expression *complément du nom* (v. 1986 : 562-581).

Dubois (v. 1991) traite de la fonction *adjoint de nom* que la *Nova gramática do português contemporâneo* appelle « *adjunto adnominal* ». Cette grammaire mentionne qu'il s'agit d'une étiquette donnée à un élément à valeur d'adjectif qui a pour fonction de spécifier ou de délimiter le signifié du substantif. Elle précise également que cet élément a un caractère accessoire. Ainsi, contrairement au *complément du nom*, l'*adjoint de nom* n'est pas indispensable à la compréhension du segment, même s'il fournit des nouvelles informations (v. Cunha et Cintra, 1985 : 145).

Dans le but de mieux systématiser l'analyse de ces deux fonctions syntaxiques pour notre recherche, nous distinguons les fonctions syntaxiques *complément du nom* et *adjoint de nom* selon des critères sémantico-structuraux. Cela signifie concrètement qu'un élément va recevoir la fonction *complément du nom* ou *adjoint de nom* en tenant compte de deux critères :

a) le type de base de l'UTC;
b) le fait que l'argument est introduit ou non par une préposition.

Il faut signaler que l'attribution de ces critères est toujours fondée sur la sémantique. Autrement dit, même si plusieurs critères peuvent être appliqués, c'est principalement à partir de critères de nature sémantique qui existent entre la base de l'UTC et son argument que l'on peut faire une distinction entre *complément du nom* ou *adjoint de nom*.

Ainsi, quand il s'agit d'une base déverbale suivie d'un syntagme prépositionnel, l'argument a la fonction syntaxique de *complément du nom*. Le *complément du nom*, comme le mot le dit, complète le signifié de la base avec qui il a une relation. Prenons par exemple l'UTC *agregado de células* (agrégat de cellules). La base *agregado* (agrégat) est une base déverbale (verbe sous-jacent : *agregar*), l'argument *de células* (de cellules) est donc un *complément du nom* parce qu'il est précédé d'une préposition et qu'il apporte un complément d'information.

Lorsque la base déverbale est suivie d'un adjectif (sans préposition), il s'agit alors d'un *adjoint de nom*. C'est le cas de l'UTC *ação endógena* (action endogène), où la base *ação* (action) est une base déverbale (verbe sous-jacent : *agir*) et l'argument *endógena* (endogène) est un terme à valeur adjectivale qui a pour fonction de spécifier et de délimiter le signifié de cette base.

En ce qui a trait aux UTC de base non déverbale, nous observons que l'expansion à droite est la plupart du temps un *adjoint de nom*. Tel est le cas d'*antera madura* (anthère mature) où l'argument *madura* (mature) est un adjectif dont la fonction est d'attribuer une valeur à la base de l'UTC.

Pour l'instant, nous n'irons pas plus loin dans cette analyse, mais nous y reviendrons en détail dans le chapitre sur l'interprétation des données.

Pour ce qui est de la dernière fonction syntaxique, le *complément circonstanciel*, elle est réservée aux éléments d'une UTC qui sont des adverbes ou qui font partie de la catégorie grammaticale dite locution adverbiale[9]. On trouve cette dénomination dans Dubois (v. 1991 : 103) et dans Grevisse (v. 1986 : 497-510). Cependant, Grevisse préfère parler de *complément adverbial non essentiel* (v. 1986 : 500) justement pour mettre en évidence le caractère accessoire de cet élément.

À la lumière de la pensée fonctionnaliste, nous interprétons le *complément circonstanciel* comme ayant une fonction d'*adverbe nucléaire* ou d'*adverbe étendu* selon le rôle joué par ces éléments dans la structure d'une UTC. En effet, ces dénominations sont fondées sur une classification de l'adverbe donnée pour Guimier. Cette classification a été établie sur deux plans distincts. Sur le plan syntaxique, Guimier différencie l'adverbe intra-prédicatif et l'adverbe extra-prédicatif (v. 1996 : 5). Le premier « [...] est intégré dans la phrase; en particulier, il n'est pas séparé intonativement du verbe et constitue avec lui le prédicat » (*ibid.*). L'adverbe extra-prédicatif, lui, « [...] n'est pas intégré dans la phrase. Il ne porte d'ailleurs pas sur le verbe mais sur l'intégrité de la phrase, qu'il modalise de l'extérieur [...] » (*ibid.*).

Sur le plan de la sémantique, Guimier classe l'adverbe comme étant endophrastique ou exophrastique. Les adverbes endophrastiques sont « [...] d'un point de vue sémantique, des constituants internes à la phrase, qui affectent le contenu même de l'élément sur lequel ils portent et, ce faisant, participent à la construction du sens référentiel de la phrase » (*id.* : 6). Pour leur part, les adverbes exophrastiques sont « des constituants externes à la phrase,

[9] Nous considérons la locution adverbiale comme une catégorie grammaticale puisque dans la grammaire de la langue portugaise elle est classée comme une sous-catégorie de la catégorie adverbe. En effet, dans le cadre de notre recherche nous allons voir que la dénomination locution adverbiale est réservée aux expressions latines du type *in vivo*, *in vitro*, *in situ*, etc. Qui plus est, dans *Le Nouveau Petit Robert* ces expressions sont également classées comme appartenant aux locutions adverbiales (v. 1995).

en ce sens qu'ils ne participent pas à la construction de son sens référentiel, mais représentent des traces de l'intervention du locuteur, qui commente tout ou partie de son énoncé ou de l'acte qui le produit » (*ibid.*).

Nous retiendrons les sens généraux de la classification de Guimier en l'adaptant à nos données. Si nous faisons un parallèle avec la structure syntaxique des UTC, nous pouvons dire que puisqu'un adverbe a un rapport avec un verbe dans le contexte d'analyse de la phrase, il peut également avoir un rapport soit avec un adjectif (dans son rôle d'argument) soit avec l'UTC entière. Lorsque le rapport se fait avec l'adjectif, l'adverbe est appelé *adverbe nucléaire*. Quant à l'adverbe qui a un rapport avec l'UTC prise comme un tout, il est appelé *adverbe étendu*. Dans les règles de formation, les *adverbes étendus* sont analysés comme des satellites puisqu'ils ont les mêmes caractéristiques relationnelles que ceux-ci. On remarque d'ailleurs qu'ils se rapprochent plus des locutions adverbiales que des adverbes simples.

Étant donné que les *adverbes nucléaires* ont un rapport avec l'argument de l'UTC, ils vont toujours être présents dans la prédication nucléaire. C'est pour cette raison qu'ils ont reçu l'étiquette *d'adverbes nucléaires*.

Nous présentons ci-dessous quelques exemples de la classification proposée. Vu que la syntaxe relève de la sémantique de l'UTC, notre classification est autant d'ordre syntaxique que d'ordre sémantique.

Voici quelques exemples d'UTC qui comportent, au plan de la syntaxe, *un adverbe nucléaire* :

- *cloroplasto* ***fotossinteticamente*** *competente* (chloroplaste photosynthétiquement compétent);
- *DNA* ***altamente*** *repetitivo* (ADN hautement répétitif);
- *espécie* ***altamente*** *endógama* (espèce hautement endogame);

- *espécie* ***sexualmente*** *compatível* (espèce sexuellement compatible);
- *espécie* ***sexualmente*** *incompatível* (espèce sexuellement incompatible);
- *indivíduo* ***genotipicamente*** *idêntico* (individu génotypiquement identique);
- *mRNA* ***altamente*** *purificado* (ARNm hautement purifié);
- *planta* ***genotipicamente*** *uniforme* (plante génotypiquement uniforme);
- *tecido* ***altamente*** *diferenciado* (tissu hautement différencié).

Nous citons également quelques cas d'*adverbes étendus* (ou, plus précisément, de *locutions adverbiales*) :

- *aclimatação de plantas* ***no solo*** (acclimatation de plantes au sol);
- *antera* ***in vivo*** (anthère in vivo);
- *cultura* ***in vitro*** (culture in vitro);
- *hibridação* ***in situ*** (hybridation in situ);
- *indução floral* ***in vitro*** (induction florale in vitro);
- *propagação clonal* ***in vitro*** (propagation clonale in vitro).

Mentionnons que ces adverbes peuvent être spécifiques au domaine ou non. Dans le premier cas, il s'agit d'*adverbes de domaine*. Cette dénomination a été suggérée par Guimier qui fait remarquer qu'il s'agit d'une « étiquette purement sémantique » (1996 : 141-142). Voyons exactement comment Guimier définit l'*adverbe de domaine* :

> Les adverbes de domaine ont pour effet de restreindre la validité de l'énoncé à un champ référentiel particulier. Ce sont à proprement parler des adverbes de cadrage notionnel. Les linguistes en font une classe qui comporte une liste ouverte d'items adverbiaux, dont de très nombreux adverbes en *-ment* (*anatomiquement*, *commercialement*, *géographiquement*, *logiquement*, *moralement*, *psychologiquement*, *syndicalement*, etc.) et les locutions prépositionnelles (*sur le plan politique*, *en matière de sexe*, *en principe*, etc.) (*id.* : 142).

Nous avons repris quelques-uns des exemples mentionnés précédemment pour illustrer les *adverbes de domaine* :

- *espécie* ***sexualmente*** *compatível* (espèce sexuellement compatible);
- *espécie* ***sexualmente*** *incompatível* (espèce sexuellement incompatible);
- *indivíduo* ***genotipicamente*** *idêntico* (individu génotypiquement identique);
- *planta* ***genotipicamente*** *uniforme* (plante génotypiquement uniforme).

Comme le montrent les exemples (v. *supra*, p. 36), les *adverbes de domaine* peuvent aussi avoir la forme d'une locution adverbiale comme *in situ*, *in vitro*, *in vivo*. Ces locutions apparaissent dans plusieurs domaines, dont la biotechnologie des végétaux.

Finalement, pour ce qui est de la fonction pragmatique, nous utilisons la même terminologie que Dik, à savoir: *topique* et *focus*. Ces deux étiquettes sont définies par Dik comme un résultat de l'organisation d'une situation de communication entre un émetteur et un récepteur. Ainsi, selon Dik, « Topic and Focus are used to capture the organisation that Speakers impose on predications with respect to the pragmatic information of Speaker and Addressee with a certain communicative exchange » (1978 : 132). Dik traite uniquement de l'acte de communication en langue générale. Toutefois, on retrouve une situation analogue en langues de spécialité, mais celles-ci gardent leurs propres particularités. Ainsi, si nous prenons en considération que le spécialiste recherche habituellement une efficacité maximale dans une situation de communication, nous pouvons dire que ce comportement sera exprimé dans la terminologie utilisée par ce même spécialiste et qu'il pourra être analysé par les fonctions pragmatiques.

Nous avons adapté les définitions de ces étiquettes de la manière suivante :

a) *Topique* : Le *topique* est attribué à la base de l'UTC à laquelle tous les arguments et satellites font référence afin de préciser l'objet de communication en décrivant, en expliquant ou en situant cette base.

b) *Focus* : Le *focus* est attribué aux éléments de l'UTC qui spécifient le statut informationnel de l'élément *topique*. Autrement dit, le *focus* ponctue l'information du *topique*.

Prenons par exemple l'UTC *célula hospedeira receptora* (cellule hôte receveuse). Les fonctions pragmatiques sont :

- *topique* = *célula* (cellule);
- *focus* = *hospedeira* (hôte);
- *focus* =*receptora* (receveuse).

La base *célula* (cellule) est considérée comme le *topique* parce que c'est avec elle que les deux autres arguments ont un rapport. Ainsi, *célula* (cellule) est l'élément central de la communication à partir duquel les autres éléments de l'UTC vont commenter, souligner, délimiter un ou plusieurs aspects pertinents.

Quant aux éléments considérés comme des *focus* tels que *hospedeira* (hôte) et *receptora* (receveuse), ils précisent le *topique* de la communication. Cela signifie que le spécialiste, quand il fait référence à *célula hospedeira receptora* (cellule hôte receveuse), le fait, entre autres, pour dire qu'il ne s'agit pas d'un autre type de cellule que la cellule qui agit à la fois comme hôte et comme récepteur.

Ces fonctions s'analysent sur le plan de la communication spécialisée de la terminologie. L'UTC est donc considérée comme une unité de communication entre les spécialistes. En tant que tel, elle a une valeur d'instrument de communication utile dans le milieu scientifique et technique. Faisant partie de la communication dans les langues de spécialités, les UTC utilisent des mécanismes linguistiques pour atteindre cet objectif d'efficacité communicative. Ces mécanismes sont régis par des règles sémantiques, syntaxiques et pragmatiques. Nous croyons que c'est en analysant ces règles (ici

interprétées comme des fonctions) que nous atteindrons notre but qui est de déterminer le comportement linguistique des UTC en langue portugaise.

1.5 Conclusion

La terminologie d'une langue de spécialité sert à nommer des notions propres à la communication d'expériences entre les spécialistes d'un domaine. Dans ce contexte, la langue est interprétée comme un instrument social de communication et le terme est considéré comme le moyen capital utilisé dans ce processus. Toutefois, les notions à transmettre ne sont pas toujours simples. Dans les textes spécialisés, les unités complexes sont souvent utilisées pour représenter des notions élaborées, c'est-à-dire des notions qui possèdent une accumulation de traits caractériels qui leur confèrent un sens précis. Ces unités terminologiques complexes emmagasinent une charge d'informations sémantiques, syntaxiques et pragmatiques qui détermine le comportement linguistique d'un domaine. À partir d'une théorie de base fonctionnaliste, ces informations sont interprétées comme des fonctions conformes à des critères selon leur rôle dans la communication spécialisée.

Étant donné la nature fonctionnelle de notre analyse, nous avons choisi la grammaire de Dik comme méthode de base pour l'interprétation des données. Cette grammaire a l'avantage de rendre la systématisation des trois types de fonctions présentes dans les UTC possible, à savoir la fonction sémantique, la fonction syntaxique et la fonction pragmatique. De plus, cette dernière fournit une base solide de principes ainsi qu'une méthode pour l'organisation de données appelée *méthode de la prédication*. Finalement, les fondements d'analyse qui inspirent cette grammaire sont très utiles pour notre recherche, puisqu'ils établissent que la syntaxe est déterminée par la sémantique et que ces deux éléments doivent toujours être analysés dans un contexte pragmatique.

Comme nous l'avons constaté au cours de ce chapitre, la grammaire fonctionnelle de Dik a été adaptée à l'analyse des données de nature terminologique. Ce qui démontre la

flexibilité de la théorie choisie. D'ailleurs, l'adéquation aux divers types de langues est un des principes préconisés par Dik.

Pour toutes ces raisons, nous croyons que le cadre théorique que nous venons d'exposer fournit toute l'information nécessaire pour la description et l'analyse des mécanismes linguistiques utilisés dans la terminologie de la biotechnologie de végétaux.

2. État de la question

2.1 Introduction

Notre recherche comporte deux volets. Le premier porte sur la description et l'analyse des caractéristiques morphosyntaxiques et syntactico-sémantiques des unités terminologiques complexes relevées dans des textes en langue portugaise. Le second volet, une conséquence du premier, vise principalement à établir des critères pour la conception d'un logiciel d'extraction terminologique. Nous donnerons également un bref aperçu de quelques outils utilisés dans la chaîne terminographique.

Étant donné ce double aspect de notre recherche (l'analyse linguistique des unités complexes et l'automatisation des données terminologiques), nous avons divisé l'état de la question en deux grandes parties. La première partie est dédiée aux travaux en matière d'analyse linguistique de l'objet terminologique. Puisque notre recherche est fondée sur des méthodes d'analyse morphosyntaxique et syntactico-sémantique dans une perspective fonctionnaliste de la langue, nous allons nous concentrer, dans cette section, sur les principales recherches en terminologie qui ont suivi une démarche fonctionnelle. Nous présenterons, dans la deuxième partie, les différents aspects de l'automatisation de la chaîne terminographique en tenant compte plus précisément des logiciels d'extraction terminologique qui reflètent la perspective fonctionnaliste.

2.2 Fonctionnalisme dans les études terminologiques

Étant donné qu'une banque de terminologie doit refléter l'usage dans les domaines de spécialité, il est normal qu'elle comporte plus d'unités terminologiques complexes que simples. L'importance de ce type d'unités se fait sentir même dans les dictionnaires de langue. À ce propos, Boulanger souligne qu' « Une recherche rapide et partielle au sujet des différentes catégories d'entrées dans cinq[10] dictionnaires de langue récents montre que le syntagme[11] occupe environ 1% de la nomenclature totale » (1989 : 516). Effectivement, selon Boulanger, « [...] les répertoires généraux ne se privent pas d'importer des vocabulaires thématiques complets ou partiels pour les ranger auprès des unités courantes figurant à la nomenclature. La part des LSP est importante dans les dictionnaires d'usage » (1995 : 83).

Cette présence, somme toute considérable, des UTC dans des ouvrages spécialisés et généraux est d'ailleurs une conséquence de la régularité de ce type de composants dans des textes scientifiques et techniques. En réalité, l'emploi des syntagmes terminologiques est le moyen utilisé pour adapter les structures cognitives des langues de spécialité au discours terminologique.

Étant donné l'importance de cet élément ainsi que l'approche retenue pour notre recherche, nous allons faire le point sur les principaux travaux réalisés sur les unités terminologiques complexes dans une perspective fonctionnaliste.

Il faut également préciser que ce ne sont pas nécessairement tous les travaux cités dans cet état de la question qui s'intéressent à l'extraction automatique de données. Nous

[10] Les dictionnaires sont : *Dictionnaire du français plus*, *Le Grand Robert de la langue française*, *Le Petit Robert 1*, et deux éditions du *Petit Larousse illustré*.

[11] La recherche se restreint au comportement de la séquence *CHE-* dans les dictionnaires de langue (v. Boulanger 1989 : 364-366).

les mentionnons quand même puisqu'ils constituent une importante source d'information dont nous ne pouvons nous passer.

2.2.1 Description et analyse des unités terminologiques complexes

Vu qu'une partie de la thèse traite de la conception de logiciels d'extraction terminologique, il est essentiel d'obtenir un consensus sur ce que nous entendons par unité terminologique complexe. De prime abord, nous pouvons penser que la littérature va nous fournir une définition unique et précise de ce concept. Pourtant, c'est loin d'être le cas. Dans ce domaine, les réflexions sont multiples et varient beaucoup selon l'approche choisie. Il faut donc les connaître pour arriver à privilégier celle qui convient le plus à nos besoins.

La majorité des recherches porte sur les rapports internes et externes des UTC. Par rapports externes, nous entendons le comportement de l'UTC dans la phrase. Les rapports internes sont pour leur part des dépendances entre les constituants d'une UTC. Même si nous croyons que la création de termes n'est pas toujours arbitraire, elle prend parfois des chemins imprévisibles. Afin de déterminer les rapports établis entre chaque composant d'une UTC, plusieurs auteurs se sont penchés sur des recherches fondées sur la syntaxe, la sémantique ou sur une perspective syntactico-sémantique. Ces recherches, menées sur un large éventail de sujets, sont importantes et permettent de mieux connaître les différentes caractéristiques des UTC, ce qui rejoint notre objectif.

Dans la tradition fonctionnaliste, Martinet nomme l'UTC *synthème* et la définit comme « un signe linguistique que la commutation révèle comme résultant de la combinaison de plusieurs signes minima, mais qui se comporte vis-à-vis des autres monèmes de la chaîne comme un monème unique » (1985 : 37). Toutefois, il y a des différences entre cette appellation créée par Martinet et celles d'autres auteurs. Selon Gross, Martinet « classe parmi les synthèmes les mots dérivés (*désirable, refaire*), structure qui relève, pour la plupart des autres auteurs, de la dérivation et non pas de la composition »

(1996 : 5). De plus, Gross considère que la définition de *synthème* de Martinet, soit « séquence formée de plusieurs monènes lexicaux fonctionnant comme une unité syntaxique minimale » ne prend pas en considération la question sémantique (v. *ibid*.). En effet, pour Martinet, un *synthème* se défini principalement par rapport à son comportement syntaxique (v. 1985 : 36 et 37).

Benveniste avait déjà proposé le nom *synapsie* pour désigner un type particulier de composition qui, d'après lui, se produit principalement dans le domaine technique. En relevant le caractère spécial de ce type de composition, il mentionne qu' « il consiste en un groupe entier de lexèmes, reliés par divers procédés, et formant une désignation constante et spécifique.[...] Le fait nouveau et important est qu'il prend aujourd'hui une extension considérable et qu'il est appelé à une productivité indéfinie : il est et sera la formation de base dans les nomenclatures techniques » (1974 : 172). Comme l'affirme Gross, la définition de *synapsie* de Benveniste est « habituellement [celle] du mot composé. Seulement, il se trouve que Benveniste [...] oppose la *synapsie* (*machine à coudre*) au mot composé (*timbre-poste*, *marteau-pilon*) et aux dérivés (*ferblanterie*) par différents critères [...] » (1996 : 5).

Afin de caractériser la *synapsie* en langue française, Benveniste propose sept traits principaux, à savoir :

- la nature syntaxique (non morphologique) de la liaison entre les membres;
- l'emploi de joncteurs à cet effet, notamment *de* et *à*;
- l'ordre déterminé + déterminant des membres;
- leur forme lexicale pleine, et le choix libre de tout substantif ou adjectif;
- l'absence d'article devant le déterminant;
- la possibilité d'expansion pour l'un ou l'autre membre;
- le caractère unique et constant du signifié (1974 : 172-173).

Comme le signale Daille, « Cette définition est intéressante, car c'est la première qui prend en compte l'extrême productivité des formes N_1 PREP N_2 et N ADJ dans les nomenclatures techniques » (1994a : 37). L'auteure ajoute également que ces critères « sont à la fois d'ordre sémantique et d'ordre syntaxique » (*ibid*.).

Guilbert est du même avis que Benveniste quand il affirme que « La démarche adoptée par E. Benveniste permet [...] d'intégrer [...] [les] variations syntaxiques dans le processus de formation des unités syntagmatiques, puisque le principe essentiel de leur définition est la liaison syntaxique entre les composants [...] » (1975 : 253).

Guilbert considère que la *synapsie* est un type de dérivation syntagmatique. Il explique d'ailleurs que par cette désignation « il faut entendre la formation d'une séquence de signes disjoints constituants une unité de signification permanente, une unité de lexique apte à fonctionner dans la phrase selon les mêmes relations syntagmatiques et paradigmatiques que l'unité appelée "*mot*" » (1970 : 116).

Pour ce qui est du comportement syntaxique interne des UTC, Benveniste avait déjà démontré que « La composition nominale est une micro-syntaxe » (1974 : 145). Pour sa part, Kocourek fait ressortir que les relations syntaxiques entre les composants d'une unité complexe obéissent à un ordre hiérarchique. En s'appuyant sur le concept de formation binaire de Guilbert, Kocourek démontre, à l'aide de plusieurs exemples, que la structure hiérarchique est établie par paires de deux mots et qu'elle est progressive, c'est-à-dire que « Sometimes the modifier itself may be further expanded » (1979 : 130).

Du point de vue de l'axe syntagmatique, cette hiérarchie peut donner lieu à des segments très longs et entraîner une perte de cohésion entre les éléments. Dans une perspective textuelle, « les déterminations successives du syntagme s'effectuent selon la linéarité de la phrase, c'est-à-dire de la gauche vers la droite [...] Le syntagme lexical est en effet une phrase réduite par une série d'effacements récursifs [...] » (Auger, 1978 : 15-16). Auger a également constaté que « plus longue est la description (et par conséquent le

syntagme lexical), plus elle s'approche de la paraphrase synonymique de la définition et donc du discours métalinguistique » (*id*. : 13). L'exemple *palan électrique à engrenage à chaîne à maillons à commande par boîte pendante à boutons-poussoirs* (v. Boulanger, 1985 : 24) illustre bien ce phénomène. Nous voyons donc qu'il est difficile de déterminer la longueur maximale d'une UTC.

Comme le fait remarquer Kocourek, les rapports syntaxiques ne sont pas toujours clairs et peuvent causer des ambiguïtés (v. 1979 : 131). Afin de lever ces ambiguïtés, cet auteur propose d'avoir recours à « certain underlying constructions revealing further the nature of the semantic relation between the constituents of the phrase » (*ibid*).

L'analyse de la composante sémantique des UTC s'avère particulièrement intéressante parce qu'elle complète les analyses syntaxiques dont il a été question auparavant. Nous allons d'ailleurs montrer un peu plus loin que les UTC peuvent également être analysées selon une perspective sémantico-conceptuelle. Ainsi, du point de vue du système conceptuel, les termes complexes sont définis comme des représentants des concepts appartenant à un domaine. Chaque concept est défini par rapport à un autre, formant ainsi un réseau conceptuel. Certains concepts sont plus spécifiques, et parfois l'utilisation d'un terme simple ne suffit pas pour les représenter. Il faut donc faire appel à une méthode systématique de désignation qui est « based on consistency in the creation of terms by choosing qualifiers indicative of some property or other essential characteristic of a concept. Regular patterns of compound terms could in theory be developed with one, two, three and more elements to indicate hierarchical dependencies » (Sager, 1990 : 64).

Par ailleurs, les recherches de type fonctionnaliste considèrent qu'« [...] une langue de spécialité est une langue fonctionnelle, dont la plus petite unité est le terme[12]; le système des termes représente le système de notions. C'est le principe de dénomination syntagmatique qui permet l'installation de cette relation de représentation entre système de notions et système de termes » (Bourigault, 1993 : 115).

La dépendance hiérarchique est établie pour rendre le réseau conceptuel d'un domaine transparent et elle est réalisée concrètement par des rapports sémantiques entre les composants d'une UTC.

Toutefois, les rapports sémantiques qui existent entre les composants des unités complexes sont parfois obscurs. À ce sujet, Habert et Jacquemin révèlent qu'il existe deux pôles d'analyse sémantique des composés.

> Pour certains, les noms dit composés sont « opaques », leur sens n'est pas calculable à partir du sens de leurs composants. Pour d'autres, à l'opposé, la sémantique de ces séquences est globalement compositionnelle, même si le poids de l'histoire, en particulier, rend parfois difficile la perception des régularités. Pour la plupart des travaux cependant, les « noms composés » sont disposés sur un continuum qui va de l'opacité totale à la compositionnalité (1993 : 13-14).

Cette opacité est appelée par Gross *opacité sémantique*. Par cette dénomination, Gross met en évidence le phénomène de la non-transparence du sens de chaque composant d'une unité complexe. Il donne comme exemple le terme *clé anglaise* qui « n'est pas une clé que l'on a fabriquée en Angleterre, mais une variété de clés, une clé d'un type particulier » (1996 : 11). L'auteur ajoute que l'opacité sémantique « est un phénomène scalaire : elle peut être totale (*la clé des champs*), partielle (*clé anglaise*) ou inexistante (*clé neuve*) » (*ibid*.).

Gross traite également du phénomène de la restriction syntaxique qui, avec l'opacité, empêche l'analyse par les propriétés transformationnelles des éléments. La restriction syntaxique est un type de figement syntaxique qui fait que l'unité complexe est toujours utilisée dans son ensemble et dans le même ordre. Gross donne l'exemple de *cordon(-)bleu* (bonne cuisinière) qui n'accepte pas les transformations suivantes :

---

[12] En linguistique générale, la plus petite unité significative est plutôt le morphème (note personnelle).

*le bleu de ce cordon

*un cordon très bleu

*un cordon particulièrement bleu

*ce cordon est bleu (v. *id*. : 12-13)

Cependant, comme le font remarquer Habert et Jacquemin, la transparence sémantique peut être très évidente au moment de l'apparition d'un mot et devenir opaque au cours de l'histoire, ce qui ne veut pas dire pour autant que la suite ne peut pas être décomposée en unités de sens individuelles. Pour illustrer leur propos, les auteurs citent l'exemple d'*arc de triomphe* de Grevisse (v. Habert et Jacquemin, 1993 : 14).

Dans une perspective d'articulation entre le sens et la référence des unités lexicales, Corbin et Temple distinguent trois niveaux de catégorisation :

a) les catégories sémantiques [qui] sont des catégories linguistiques définies par l'ensemble des propriétés sémantiques constituant les sens d'un mot et sont caractérisables par leur intension;
b) les catégories préréférentielles [qui] sont le résultat des découpages conceptuels que les propriétés sémantiques permettent d'opérer et de dénommer et sont caractérisables en intension et extension;
c) les catégories référentielles [qui] sont les catégories préréférentielles adaptées en fonction de notre appréhension pragmatique et culturelle des choses et forment un sous-ensemble des catégories conceptuelles (1994 : 5-6).

Ainsi, Corbin et Temple accordent une importance majeure au sens lexical, présupposé de base de leur théorie. Cette théorie est fondée sur la construction du sens à partir de l'analyse dérivationnelle et sémantique du mot. En mettant en relief la morphologie, les auteurs soutiennent que « le sens des mots construits est étroitement associé à leur structure morphologique, autrement dit qu'il existe un sens prédictible à partir de la façon dont un mot est construit, qui définit l'identité sémantique de chaque mot

particulier » (*id.* : 8-9). En d'autres mots, le auteurs croient en l'existence « des catégories sémantiques associées aux mots » (*id.* : 7).

Cette théorie du cadre morphologique est d'ailleurs appliquée par David dans une étude sur la formation *N +N* en français. En réalité, l'auteure met en jeu une analyse tant syntaxique que morphologique. À l'aide d'observations morphologiques, elle « [...] définit une règle de construction de mots (RCM) comme une règle qui associe à une structure (ou à plusieurs) un sens compositionnel par rapport à celle(s)-ci » (1993 : 63). Par le biais de l'analyse syntaxique, David prouve que « l'analyse linguistique d'une phrase n'est pas entièrement déductible des unités lexicales qui la constituent. On distingue dans ce modèle un ensemble de positions, c'est-à-dire un ensemble de "points" hiérarchiquement organisés entre eux, formant une configuration, et une relation entre les positions et les termes » (*ibid.*).

Revenons maintenant aux recherches qui ont privilégié une perspective onomasiologique. Habert et Jacquemin classent les diverses recherches menées dans cette perpective en quatre types :

- les typologies des relations sémantiques régissant les composés;
- l'analyse sémantique des composés à tête déverbale;
- les prépositions « incolores » comme « leviers » pour l'interprétation;
- l'élargissement de la notion de prédicativité nominale à des constituants autres que des déverbaux (1993 : 16-18).

Il est évident que travailler avec la sémantique de mots n'est pas une tâche facile. Abraham souligne la problématique en relation avec la définition d'une typologie sémantique influencée par le fonctionnement syntaxique d'un terme. « D'un côté, le typage sémantique a priori dépend de la perception physique ou culturelle des propriétés attachées à une entité, mais d'un autre côté, le type d'une entité peut changer lorsqu'elle est impliquée dans des relations avec d'autres entités, i.e., lorsqu'elle instancie des schèmes

dans un contexte » (1994 : 134). L'auteur cite l'exemple de *voiture* qui, selon son entourage contextuel, peut, entre autres, être un lieu (p. ex. *Pierre est dans la voiture*), une collection (p. ex. *un paquet de voitures*). Ainsi faisant, il démontre que « des procédés systématiques (catégorisateurs) permettent de changer les types premiers d'une entité » (*id.* : 136).

Afin de rendre les rapports entre les composants des UTC plus clairs, Sager et Kageura (v.1994/1995) proposent une méthode intéressante pour l'établissement des classes conceptuelles qui peuvent servir à l'analyse des rapports entre les concepts représentés par des UTC. Nous reproduisons leur classification dans le tableau qui suit.

Tableau 2.1 *— Classification des concepts*

Entity concepts	material entity, non-material entity
Activities concepts	action, transfer, state change, production.
Relations concepts	subclass of concepts which identify the relationships that exist or are established between entities, between activities, between properties and among each of these classes in combination.
Property concepts	subclass of concepts which are used (a) to identify and describe differences between concepts and activities and (b) to identify features or characteristics of entity concepts, activity concepts or relation concepts.

Même si ces auteurs n'ont pas abordé directement la question de la délimitation terminologique, nous pensons que ce type de recherche fondamentale peut être très utile pour mieux comprendre la structure interne des UTC.

Dans la même voie sémantico-conceptuelle, nous retenons spécifiquement la recherche menée par Pugh qui examine les modes de formation des termes du domaine de l'informatique dans trois langues : anglais, français et espagnol. Pour ce faire, elle a établi neuf classes de référents conceptuels et dix-huit classes relationnelles pour représenter les rapports conceptuels établis entre les éléments des unités complexes (v. 1983). Cette auteure, comme plusieurs autres, en arrive à la conclusion que « naming in special language is indeed less arbitrary and more systematic than general language designation and is

strongly influenced by the need to reflect and reinforce the systematized conceptual structure of a particular subject field » (*id*. : 399).

Dans une autre étude, moins récente mais tout de même significative, Dowing traite de la création et de l'usage du *N + N* en anglais. Cet article est particulièrement intéressant parce que l'auteure se penche non seulement sur la fonction sémantique mais aussi sur la fonction pragmatique des composants de l'UTC. Dowing mentionne que

> [...] constraints on N + N compounds in English cannot be characterized in terms of absolute limitations on the semantic or syntactic structures from which they are derived. Indeed, because of the important differences in the functions served by compounds, as opposed to the sentential structures which more or less accurately paraphrase them, attempts to characterize compounds as derived from a limited set of such structures can only be considered misguided. A paraphrase relationship need not imply a derivational one (1977 : 840-841).

Il faut noter également l'étude de Levi qui propose d'analyser les propriétés syntaxiques et sémantiques des unités complexes. Dans son analyse, l'auteure propose un ensemble de structures sémantiques associées à la formation des unités complexes. En outre, on remarque l'importance de la perspective pragmatique dans une étude de cette nature. À ce sujet, elle précise que

> In the course of this exploration, it has become clear that a complete description of the role of complex nominals in natural language must include not only the kind of syntactic and semantic facts that formal derivations can account for, but also a description of the broader semantic and pragmatic principles that influence the ways in which both speakers and hearers manipulate the formal regularities of Cns[13] in actual discourse (1978 : 237).

Les réflexions sur la formation des unités complexes dans une perspective sémantique sont également reprises dans la recherche menée par Leonard (1984). En effet, cette dernière établit une liste des caractéristiques sémantiques utilisées par l'ordinateur

[13] Cns signifie *complex nominals*.

pour interpréter des phrases nominales (*noun sequences*) en anglais. Nous voyons ainsi que la question sémantique est fondamentale en tant qu'élément d'analyse des rapports entre les arguments d'une structure prédicative des unités terminologiques complexes.

Toujours en ce qui concerne les travaux s'intéressant aux classes et aux rapports conceptuels entre les éléments des UTC, citons les études de L'Homme et Gagnon (à paraître) et de L'Homme (1996, 1996a et 1996b) sur la formation des UTC en français et le choix prépositionnel. Dans l'un des ces articles, L'Homme démontre que le choix prépositionnel, dans les termes complexes du type *N prép. N*, est régi par l'organisation conceptuelle que le terme complexe tente de refléter (v. 1996).

Dans un domaine plus vaste, celui de la phraséologie, L'Homme illustre également qu'il est possible d'appliquer la méthode d'analyse des rapports conceptuels à la traduction et d'obtenir de bons résultats (L'Homme, 1995 et L'Homme et Gemme, 1997). Même si cette étude dépasse l'objet de notre recherche (les UTC) et que l'application à la traduction n'est pas notre but, les idées de base demeurent intéressantes. D'ailleurs, dans la conclusion de son article, l'auteure signale que la stratégie proposée

> [...] entails a better understanding of the function of varied lexical units in special languages. The long-standing dichotomy between specialized and non specialized units needs to be refined further. Word-combination problems illustrate only one aspect of the issue. Studies of lexical units and their function will provide for a more accurate description of special languages. They might even throw a new light on a specific item, i.e., the complex term (1995 : 160).

Toutes les études que nous venons de mentionner tentent d'expliquer le fonctionnement du processus de formation des unités complexes en terminologie. Elles vont de l'analyse des rapports syntaxiques ou sémantiques internes de l'unité complexe jusqu'à la réflexion sur l'unité prise comme un tout et son fonctionnement dans l'axe syntagmatique de la phrase. Ces deux aspects sont très importants pour déterminer la

longueur des UTC dans un contexte, phénomène essentiel pour la conception d'un logiciel d'extraction de données.

Nous entendons par longueur, ce qui se trouve entre le début et la fin d'une séquence contrainte qui forme une unité terminologique complexe. En d'autres mots, la longueur est l'étendue ou la délimitation d'une UTC. Pour fixer cette délimitation, il faut avoir des critères fondés sur le comportement des UTC. À ce sujet, Boulanger distingue les critères externes et internes nécessaires pour réussir à isoler les termes d'un contexte donné. Selon l'auteur, « par externe, nous entendons la possibilité de tracer des limites à gauche et à droite du groupement des unités reconnu comme syntagme. Par interne, nous entendons la possibilité ou l'impossibilité d'introduire des éléments lexicaux étrangers entre les éléments constituants du syntagme » (1978 : 173).

Faulstich propose d'adopter les principes de la linguistique fonctionnelle pour déterminer de la longueur des UTC qui se trouvent dans les textes de langues de spécialité. Selon Faulstich, ces principes peuvent considérablement aider le terminologue à analyser le fonctionnement des termes en accordant un statut primordial à la syntaxe et à la sémantique. Ces principes visent à :

a) attribuer à l'UTC le rôle de prédicateur sémantique;
b) adopter le critère de la prédication syntactico-sémantique pour la délimitation de l'UTC (v. 1995 : 3-4).

Nous croyons que les critères mentionnés ci-dessus ont un rapport avec le phénomène du *figement*, également appelé, par d'autres auteurs, *lexicalisation des unités complexes*. En effet, la notion de *figement* peut influencer la conception de ce qui est compris par UTC et, par conséquent, la manière de concevoir l'extraction automatique. Même si la lexicalisation a une influence sur la conception d'un logiciel d'extraction de données terminologiques, celle-ci n'est pas primordiale. Nous sommes d'accord avec David lorsqu'elle affirme qu'il faut faire une différence entre le mode de construction d'une unité

et sa lexicalisation. Le mode de construction appartient à la langue, tandis que la lexicalisation relève de l'usage qu'on fait de cette unité. La lexicalisation se situe donc hors de la langue. Ainsi, « [...] pour qu'une unité se lexicalise il faut qu'elle soit ...déjà construite. La lexicalisation n'est donc pas identificatoire du mode de construction des mots dits composés » (David, 1993 : 62).

Toutefois, cette façon de raisonner est loin de faire l'unanimité chez les auteurs. Nous présenterons, dans la section qui suit, les divers critères proposés dans les ouvrages qui traitent du figement. Il faut signaler que nous allons principalement mettre en évidence les critères qui pourront, d'une certaine manière, être utiles pour la conception d'un logiciel d'extraction terminologique.

2.2.2 Critères de figement

La problématique du figement débute par la notion même exprimée par ce mot. Bon nombre d'auteurs se sont penchés sur la question, et les définitions diffèrent beaucoup en fonction de l'approche choisie. La cohésion, c'est-à-dire la difficulté d'insertion et de modification d'un des composants d'une unité complexe, est toujours un élément clé dans la notion de figement. Toutefois, comme le fait remarquer Gross, il existe un degré de figement, « un continuum entre les séquences libres et celles qui sont entièrement contraintes » (1996 : 17). Selon lui, le figement peut être soit partiel, c'est-à-dire « porter sur le déterminant du nom ou sur son expansion » (*id.* : 39), soit complet, autrement dit « Aucun des deux mots ne peut faire l'objet d'une substitution ni de modifications syntaxiques [...] » (*ibid.*).

Dans ce sens, Gross propose une distinction entre ce qu'il appelle un *groupe nominal ordinaire ou libre* et un *nom composé*. Ce classement est fondé sur des degrés de figement. Nous présentons ci-dessous un tableau qui résume les principales caractéristiques définies par Gross (*id.* : 27-28).

***Tableau 2.2** — Principales différences entre groupe nominal ordinaire ou libre et nom composé*

Groupe nominal ordinaire ou libre	**Nom composé**
Moins figé	Plus figé
En regard de sa distribution interne, il a plus de liberté de fonctionnement	En regard de sa distribution interne, il a moins de liberté de fonctionnement
A une structure et des relations internes qui permettent des reformulations	Fonctionne comme un seul bloc du point de vue de ses relations avec le reste de la phrase
Constitué d'un substantif-tête et de sa détermination	Le deuxième élément ne joue pas un rôle de modifieur

Lerat propose une échelle de figement et « l'applique à l'élément syntaxiquement le plus dépendant (par opposition à la tête, au régissant) » (1994 : 152). Son étude, qui porte sur les séquences du type *N de N*, fait ressortir les degrés suivants :

- degré 0 : distribution catégorielle fondée sur la non-prévisibilité de la tête par rapport à l'élément dépendant;
- degré 1 : distribution syntactico-sémantique fondée sur la non-pertinence de l'application des traits tels que *concret*, *abstrait*, *humain* et *animé*;
- degré 2 : distribution sémantique fondée sur le caractère aléatoire de la sémantique de traits comme moyen d'établir une relation d'hyperonymie;
- degré 3 : distribution lexicale fondée sur la non-occurrence d'un synonyme, d'un hyperonyme ou d'un hyponyme par rapport au dépendant (v. *id*. : 154-155).

On peut également supposer que certains facteurs extra-linguistiques sont des critères pour le figement. Ainsi, comme le soulignent Habert et Jacquemin, « Le fait de considérer une séquence comme une unité codée est un jugement situé dans le temps et la durée [...]. C'est un jugement dépendant souvent aussi d'un type de discours déterminé » (1993 : 9).

Kocourek divise les critères de figement en trois types, à savoir :

- les critères syntaxiques;
- les critères sémantiques;
- les critères d'usage (v. 1991 : 138-151).

Pour ce qui est des critères syntaxiques, Kocourek mentionne qu'un « groupe de mots n'est pas n'importe quelle suite de mots, c'est une séquence à structure syntaxique » (*id*. 139). Ces propos vont dans le même sens que ce que Benveniste constate au sujet des matrices syntaxiques des UTC. Selon Benveniste, « les termes de base à l'aide desquels se constitue une nomenclature tendent à devenir explicites et à constituer à leur tour des combinaisons explicites, au moyen d'unités lexicales indépendantes, identifiables et organisées selon des modèles syntaxiques » (1974 : 173-174).

Citons les principaux critères syntaxiques signalés par Kocourek :

- la basse fréquence des conjonctions et des déterminants, et l'inexistence des pronoms et des verbes auxiliaires;
- la faible constatation de séquences trop étendues;
- l'existence d'un sigle comme équivalent à une séquence donnée;
- le degré de cohésion interne entre les éléments (v. 1991 : 138-144).

En ce qui concerne spécifiquement la cohésion interne, Gross considère que « dans les séquences figées, l'insertion d'éléments nouveaux est très réduite » (1996 : 18). Il ajoute que cette rigidité compositionnelle est vérifiée tant pour les nominales contraintes[14] (p. ex. *col vert*[15], *col blanc*[16]) que pour les adjectifs composés (v. 1996 : 18-19). À ce sujet, Boulanger observe que « L'impossibilité d'insérer un élément lexical nouveau entre les

[14] L'auteur désigne par nominale contrainte les séquences de base nominale qui portent une restriction « sur la liberté combinatoire potentielle [...] et sur la calculabilité du sens » (Gross, 1996 : 154).
[15] Type de canard (v. *ib*. : 18).
[16] Bureaucrate (v. *ibid*.).

éléments constituants du syntagme tend à garantir à celui-ci une qualité d'autonomie lexicale, à consacrer sa stabilisation, à préparer sa lexicalisation, puis son enregistrement dans une nomenclature » (1978 : 176).

Sur le plan de la structure interne des mots, Gross souligne qu'il est essentiel de déterminer la typologie des formations syntaxiques d'une terminologie d'un domaine pour établir les figements et, par conséquent, les critères de reconnaissance automatique. Selon lui, « [...] pour reconnaître automatiquement les noms composés, il faut être en mesure d'en montrer les limites et prédire leur morphologie (formation du pluriel et éventuellement du féminin), et établir au préalable une typologie de la composition, afin de percevoir les problèmes spécifiques à chaque type. En fonction de la structure interne des noms composés, les éléments de la description varient » (1996 : 48). Ainsi, comme le précise l'auteur, « [...] la solution consiste à faire la liste de toutes les propriétés syntaxiques qui caractérisent les relations entre les éléments de ce groupe nominal » (1990 : 683).

En analysant la matrice *N + A* en français, Gross propose des critères pour tester le degré de figement de ce type de formation. Examinons deux des critères mentionnés par Gross. Le premier critère fait référence au rôle syntactico-sémantique de l'adjectif par rapport au substantif. Ainsi, selon Gross, si on peut déterminer ce rôle, c'est qu'il y a un faible degré de figement. Le deuxième critère attire l'attention sur la compositionnalité du sens. Plus le sens de la suite est compositionnel, moins il y a degré de figement. (v. 1996 : 50-52). Gross propose des critères semblables pour la suite *N de N* (v. *id*. 53-57).

À part le critère de l'existence d'une définition spécialisé, Kocourek mentionne, parmi les critères d'ordre sémantique, la substitution de l'UTC comme un tout. Lorsque la commutation est possible, on peut affirmer que l'UTC existe. Il faut, par contre, faire attention au syntagme descriptif non lexicalisé qui, normalement, ne doit pas être considéré comme une unité figée (1991 : 144-145).

Dans ce contexte de commutation, Gross inscrit la notion de « blocage des paradigmes synonymiques ». Cette expression est utilisée par l'auteur pour expliquer la possibilité de « remplacer un mot soit par un autre de la même classe sémantique soit, hypothèse plus restreinte, par un synonyme » (1996 : 17). D'après l'auteur, les expressions figées ne permettent pas de changer l'un de ces composants (v. *id*.: 17-18).

L'imprévisibilité sémantique[17] est un autre critère de cet ordre cité par Kocourek. Selon lui, l'existence de l'imprévisibilité confirme le figement. Il fournit six facteurs qui peuvent contribuer à l'imprévisibilité et qui se vérifient soit dans les mots composés ou dans les UTC :

- structure syntaxique hiérarchique complexe du syntagme, p. ex. *gardien de nuit*;
- rapports sémantiques sous-jacents entre les constituants, p. ex. *élection présidentielle*;
- ambiguïté des mots constitutifs, p. ex. *circulation sanguine* et *tempérament sanguin*;
- emploi figuré des mots constitutifs, p. ex. *sel de terre*;
- caractère exocentrique du syntagme, p. ex. *mort-aux-rats* (poison);
- emploi vraiment idiomatique (unique) d'un constituant, p. ex. *bête à bon Dieu* (coccinelle) (v. 1991 : 147-148).

Lorsqu'on analyse les critères sémantiques, il faut considérer les rapports établis entre les éléments d'une séquence donnée. Ces rapports sont dans la plupart des cas régis par la base de l'unité. D'après Gross, l'expansion peut désigner :

- un domaine (p. ex. : *vérité statistique*);
- une relation de prédicats à sujet (p. ex. : *promesse de Gascon*);
- une relation de prédicats à objet (p. ex. : *rempailleurs de chaises*)

[17] Gross l'appelle *opacité sémantique* (v. supra p. 46).

- une relation semi-prédicative (p. ex. : *ours blanc*)
- une relation partie à tout (p. ex. : *barreau de chaise*)
- une relation circonstancielle (substance, lieu, temps, conséquence) (p. ex. *ours polaire*) (v. 1996 : 44-46).

Quant au critère d'usage, Kocourek souligne la récurrence attestée d'une unité complexe qui est fondée sur la fréquence (v. 1991 : 149-150). Toutefois, comme le souligne Boulanger, « Par leur relation privilégiée avec le référent, les termes ne comportent pas d'aspect quantitatif au même sens qu'en langue générale. La rareté ou la nouveauté du terme ont autant de valeur signifiante en science et en technique que la fréquence d'apparition du mot en langue générale. Il suffit qu'il y ait un seul besoin pour que le terme apparaisse une seule fois et reste » (1978 : 180).

En guise de conclusion à cette section, nous voulons mettre en lumière, tout particulièrement, la question de la variation terminologique. Le phénomène de la variation joue sans doute un rôle fondamental qui doit être pris en compte au moment d'établir des critères de figement. Même si la variation ne constitue pas un but immédiat de notre recherche, nous avons remarqué que les descriptions linguistiques qui ne prennent pas en considération les variantes présentent, selon nous, un point de vue incomplet de la langue.

En effet, comme l'explique Daille et *al* « Terms are subject to different kinds of changes. Given a list of controlled terms, observation of technical corpora shows that variations can be roughly divided into morphological and syntactical variations » (1996 : 198). Dans leur étude, ces auteurs montrent des variations d'ordre syntaxique dans le domaine de la médecine autour du terme *epithelial cells*. Ces auteurs ont constaté quatre types de variations syntaxiques, soit l'insertion, la permutation, la coordination et la morphodérivation.

La grande majorité des travaux sur la variation sont fondés sur la socioterminologie, c'est-à-dire une perspective d'analyse de la terminologie qui privilégie le fonctionnement

de la langue dans son usage social. Analyser l'usage implique la prise en compte des variations, car l'usage n'est pas toujours stable.

À partir d'une vision fonctionnaliste de la terminologie, Faulstich présente une liste des variantes les plus communes dans les langues de spécialité. Nous reproduisons sa typologie ci-dessous :

- variantes terminologiques linguistiques : morphosyntaxiques, lexicales, phonologiques et graphiques.
- variantes terminologiques d'enregistrement : géographiques, discursives, temporelles (v. 1997 : 22-27).

Toutes ces observations sur les diverses conceptions des UTC, qui s'appliquent également au portugais, ainsi que les nombreuses réflexions sur le figement, nous portent à croire que l'organisation de la langue peut fournir des contraintes intéressantes quant à la délimitation et au sens de ces unités. Toutefois, c'est le spécialiste qui a, et qui aura toujours, le dernier mot sur le figement d'une UTC et sur le sens des ces suites complexes. Encore aujourd'hui, bien que la technologie ait connu des progrès considérables, le savoir extra-linguistique apporté par le spécialiste est une tâche qui demeure difficile à automatiser.

2.3 Informatique et terminologie

Chaque domaine scientifique et technique possède une structure conceptuelle représentée par un ensemble terminologique. Connaître et maîtriser la terminologie de façon à la rendre accessible à la communauté de spécialistes n'est pas une tâche facile. Les concepteurs de banques de terminologie le savent. Depuis leur naissance, ces véritables « trésors » électroniques accessibles au public ont vécu plusieurs difficultés qui ont exigé la créativité et la compétence des terminologues pour s'en sortir. Ces difficultés, présentes

encore aujourd'hui, se font essentiellement sentir dans trois étapes principales de la chaîne terminographique, à savoir :

a) le recensement terminologique;
b) le traitement des unités terminologiques;
c) la récupération des données terminologiques.

En ce qui concerne le recensement, même si la cueillette des données peut se faire directement auprès du spécialiste, les terminographes privilégient la collecte textuelle. Ainsi, les textes sont considérés comme des témoins de l'existence des termes dans un domaine. Toutefois, le milieu discursif apporte, lui aussi, une série de difficultés. Tel est le cas de la délimitation terminologique qui demeure encore de nos jours une activité plutôt complexe. Du côté quantitatif, l'utilisation de plus en plus fréquente des logiciels pour exécuter cette tâche rend le travail moins ardu. Drouin explique d'ailleurs que « [...] l'ordinateur est infiniment plus systématique que l'humain pour accomplir des tâches répétitives. Lancé sur un corpus de grande taille, un ordinateur peut l'analyser selon un ensemble de règles qui seront respectées du début à la fin de la procédure [...] » (1997 : 45). De plus, comme le signale Lauriston, « Given the predominance of noun phrases in technical vocabulary and the growing volume of text in machine-readable form, the automatic recognition of complex terms could accelerate appreciably terminological research » (1994 : 147). Cependant, pour arriver à concevoir un logiciel efficace pour cette étape du travail terminographique il faut prendre en compte une série de contraintes liées au comportement linguistique des termes. Dans ce contexte, Otman mentionne que « [...] l'extraction d'information de textes en langage naturel nécessite, en amont, une description linguistique des unités recherchées, description modélisable sous la forme de représentations abstraites qui pourront être traitées en machine » (1991 : 63).

Outre les problèmes relatifs au repérage textuel de la terminologie qui constituera le corpus d'une banque, il faut prendre en considération les questions concernant le traitement et la récupération du contenu linguistique et extra-linguistique de chaque terme saisi. Il faut

d'ailleurs ternir compte des besoins de récupération des données dès l'étape du traitement terminologique. Ces préoccupations sont importantes pour tous les types de banques, mais elles sont primordiales quand il s'agit des banques de terminologie fondées essentiellement sur des systèmes conceptuels. Cette nouvelle génération de banques terminologiques est née grâce à l'évolution récente qu'ont connue les recherches en informatique, plus spécifiquement dans le domaine de l'intelligence artificielle. D'un rôle passif d'enregistreurs de données, les banques contemporaines sont devenues de véritables banques de traitement et de récupération des connaissances humaines.

Dans le but de trouver des solutions pour résoudre les difficultés mentionnées précédemment, les concepteurs des banques mettent à profit plusieurs techniques du domaine de l'informatique, spécialement celles qui concernent le traitement automatique de la langue (TAL). À l'intérieur de ce vaste domaine, les terminologues se sont intéressés particulièrement aux études de l'automatisation des trois niveaux d'analyse linguistique : la morphologie, la syntaxe et la sémantique. À titre d'exemple, citons les tâches mises au point dans le cadre de l'analyse lexicale automatisée :

a) segmentation (délimitation des frontières de mots et des frontières de phrases);
b) lemmatisation (repérage du mot sous sa forme canonique);
c) étiquetage (attribution de la bonne catégorie morphosyntaxique pour une forme donnée en contexte) (v. Abeillé et Blache, 1997 : 70).

Ces techniques s'insèrent dans un domaine plus vaste, celui des industries de la langue. Abbou et *al*. désignent par ce nouveau domaine, « [...] tous les produits, techniques, activités ou services qui appellent un traitement automatique de la langue naturelle » (1987 : 30). Plus récemment, comme nous le fait remarquer Auger, au concept d'industries de la langue s'ajoute celui de la communication. On parle alors des nouvelles technologies de l'information et de la communication (v. 1998 : 5).

Parmi les secteurs d'activités des industries de la langue, on en trouve un dédié au traitement de la documentation écrite qu'on appelle terminotique. Selon Auger, la terminotique peut être considérée comme « un concept pragmatique étroitement relié à ce qu'on appelle aujourd'hui l'industrialisation de la langue, un concept générateur de produits performants de traitement de l'écrit [...] » (1989b : 485).

Pour Cabré, « La terminotique est la discipline qui s'intéresse, en général, aux relations entre l'informatique et la terminologie; et, en particulier, qui traite de l'application de l'informatique au travail terminologique »[18] (1993 : 359).

En effet, la terminotique s'insère au sein du concept des logiciels conçus pour assister le travail humain propre au traitement de la langue. C'est ainsi que selon l'Homme,

> Aujourd'hui on commence à parler de la terminotique ou, si on veut de la terminologie assistée par ordinateur dont les applications ressemblent à celles de l'informatique lexicographique : dépouillement de textes automatique, recherche automatique d'informations textuelles, découpage de contextes, lemmatisation, indexation automatique, utilisation de systèmes experts, etc. (1988 : 18)

En ce qui concerne précisément les démarches effectuées dans le cadre des systèmes experts, celles-ci ont une branche réservée aux banques de connaissances terminologiques. Dans ce domaine, les systèmes experts sont utilisés pour la résolution des problèmes liés à la structuration sémantique des notions ainsi qu'à la récupération du contenu. En effet, les systèmes experts sont « des assistants de mémoire et de capacité de calcul et de raisonnement dont la vocation est l'aide à la décision d'un usager, lui-même expert d'un domaine du savoir humain » (Otman 1989 : 64). Étant donné que la terminologie est une composante du langage humain et qu'elle traite des aspects qui concernent la sémantique de la langue, « le problème du traitement du langage naturel, et à l'intérieur de ce dernier, la

[18] Traduction personnelle du texte original rédigé en espagnol.

question du sens et de l'analyse de discours constituent deux pierres d'achoppement des travaux menés en intelligence artificielle, en particulier la branche appelée "ingénierie de la connaissance" » (*id*. 64 et 65).

Par rapport à la constitution des banques de données intelligentes, il y a encore un long chemin à parcourir puisque les systèmes experts actuels n'ont pas progressé suffisamment pour satisfaire les besoins d'une base de données dont la structure est conceptuelle. Auger remarque que

> comme solution transitoire entre la base de données « bête » et le système expert basé sur un cumul de connaissances, des progiciels de terminotique ont été développés ces dernières années qui « roulent » sur micro-ordinateur (PC compatibles) et qui permettent de constituer des bases de données terminologiques relationnelles et de les éditer et qui, de plus, sont adaptés au traitement terminologique systématique en tenant compte de la structuration notionnelle du domaine traité (1990 : 673).

Comme nous venons de l'expliquer brièvement, dans l'environnement des banques plus traditionnelles, les techniques élaborées dans le domaine de la terminotique ont également leur place dans plusieurs phases de la chaîne terminographique. Mentionnons, par exemple, l'utilisation des logiciels d'extraction terminologique et la constitution des systèmes de traitement et de récupération du contenu des bases de données terminologiques. À ce sujet, Auger précise à quels moments l'informatique peut aider les terminographes dans leurs tâches (v. 1989a : 452-453) et présente les possibilités qui existent sur le marché et dans le domaine de la recherche universitaire (v. 1994 : 47-51).

Finalement, il faut ajouter que l'utilisation des outils informatiques entraîne obligatoirement un changement des procédures standard en ce qui concerne le travail terminographique. Sager signale d'ailleurs que

> Processing systems for lexical and terminological data have been implemented using standard data processing methodologies but the result is normally far from satisfactory. This leads to the recognition that the storage

> of lexical data cannot simply be regarded as another application of traditional information processing techniques but that new techniques have to be developed for adequately representing the type of information contained in specialised dictionaries (1990 : 129).

Lauriston partage cette opinion quand il fait référence au logiciel d'extraction de données terminologiques. Selon cet auteur, « To automate, even partially, this important extraction task, it is necessary to overcome at several levels text-analysis difficulties that are not necessarily apparent to the terminologist experienced in using conventional methods » (1994 : 148).

Il est également important de souligner que quelque soit l'outil terminologique mis au point, il faut nécessairement qu'il soit basé sur des connaissances linguistiques du domaine. En d'autres mots, le terme est avant tout un signe linguistique. Il possède donc des caractéristiques morphologiques, syntaxiques et sémantiques qui devront être analysées sur un plan pragmatique en rapport avec le domaine scientifique et technique concerné. Les logiciels créés pour l'extraction terminologique doivent, à notre avis, refléter ces caractéristiques.

2.3.1 Extraction automatisée d'unités terminologiques complexes

Le découpage terminologique, soit manuel, soit semi-automatisé ou entièrement automatisé, nécessite une connaissance approfondie des caractéristiques linguistiques de la terminologie du domaine. Une terminologie dont les caractéristiques sont mal connues risque de compromettre non seulement le dépouillement, mais également de rendre le travail terminologique difficile puisque le terme est la clé d'accès à toutes sortes d'autres informations terminologiques et textuelles.

Si le découpage manuel permet au terminologue d'utiliser tout son potentiel intellectuel de réflexion sur le sujet, le découpage assisté ou automatisé, quant à lui, fournit

la rapidité et la vérification systématique de certaines tâches impossibles à effectuer pour l'être humain quand il s'agit d'un très grand ensemble de données.

Il existe aujourd'hui un grand nombre de logiciels destinés à l'extraction terminologique. Ces logiciels comportent différents degrés de complexité selon l'importance attribuée aux aspects morphologiques, syntaxiques, sémantiques et pragmatiques du terme, ce qui caractérise leur type de fonctionnement. On trouve des logiciels qui vont de l'application pure de la statistique jusqu'aux systèmes plus complexes qui utilisent des algorithmes basés sur l'analyse syntaxique du syntagme dans le texte.

De Schaetzen énumère trois types d'extracteurs de données terminologiques classés en fonction de leur puissance technologique :

- les logiciels de recherche plein-texte qui « [...] permettent de localiser des chaînes de caractères demandées par l'utilisateur, avec ou sans masque pour une ou plusieurs lettres [...] »;
- les concordanciers qui permettent l'extraction des contextes (certains requièrent d'abord un processus de lemmatisation);
- les dépisteurs de termes et de contextes qui sont des logiciels d'extraction basés notamment sur un analyseur syntaxique (1997 : 84-85).

Ces logiciels peuvent utiliser différentes techniques. Perron cite notamment

a) la constitution et l'exploitation de dictionnaires électroniques;
b) l'exploitation d'indices de fréquence;
c) l'élaboration et l'exploitation de patrons de fouille;
d) le découpage basé sur la description linguistique de la langue (v. 1989 : 4-6).

Pour ce qui est de la première technique, elle consiste à repérer des mots simples ou complexes qui existent et dans les textes et dans le dictionnaire électronique. Il s'agit d'une

technique rapide et fiable, mais qui ne peut être adoptée que si les termes susceptibles d'être relevés sont connus au départ. De plus, il serait très difficile d'avoir un dictionnaire où l'on pourrait trouver tous les termes qui existent dans les textes à exploiter, surtout quand il s'agit de textes provenant d'un large éventail de domaines, ou que ces termes sont des néologismes. À ce sujet, Otman constate que

> Ces dictionnaires, s'ils étaient constitués, allieraient le double handicap d'être fatalement incomplets et nécessairement trop volumineux pour une exploitation informatique rapide. D'autre part, la raison d'être du dépouillement automatique est de participer à la constitution et au développement de banques de données terminologiques. Si les outils de dépistage automatique se référaient à l'objet même de leur fonction, ils ne pourraient être qu'inadéquats car nous construirions un serpent se mordant la queue (1991 : 68).

En ce qui concerne la deuxième technique, elle consiste à calculer pour chacune des formes qui apparaissent dans les textes à traiter leur nombre d'occurrences. Cette méthode permet aussi de repérer les mots coocurrents, c'est-à-dire les combinaisons de mots les plus fréquentes dans le texte, ainsi que d'élaborer des algorithmes fondés sur des calculs probabilistes.

La principale difficulté de cette technique est l'élimination des hapax, soit les mots qui apparaissent une seule fois dans un texte. En terminologie, les hapax font partie d'une catégorie de termes très importante puisqu'ils peuvent représenter des néologismes potentiels. Par contre, cette méthode offre des avantages considérables comme le gain de productivité dérivé de la rapidité à présenter des résultats.

Pour ce qui est de l'élaboration et de l'exploitation de patrons de fouille, il s'agit du repérage, par ordinateur, d'une séquence de mots selon une matrice syntaxique préétablie. La rapidité est l'un des grands avantages de cette technique. Par contre, c'est une technique

qui produit beaucoup de bruit[19]. Perron explique que cette technique « ne tient pas compte de la segmentation syntaxique des phrases ou des propositions » (1989 : 5). Examinons les deux exemples suivants présentés par Perron :

- *Il transporte ses **outils d'atelier** en atelier.*
- *La mise en **relation de termes** désignant des notions apparentées...*

Le système basé sur le patron *N de N* récupère *outils d'atelier* et *relation de termes* qui, comme l'a constaté Perron, « ne peuvent, du moins dans ces contextes, constituer des unités parce que les substantifs qu'ils comportent n'entrent pas, sur le plan de la syntaxe, dans la composition du même segment » (*id.* : 9).

Il semblerait d'ailleurs, comme le mentionnent Ladouceur et Drouin, que

> l'identification automatique des termes à partir de patrons de formation syntagmatique est limitée par l'ampleur de l'intersection entre les termes et les enchaînements syntaxiques possédant une structure semblable (*panier de pommes, kilo de pommes*). Des plus, la rigidité des patrons de formation (*N1 de N2*) ne permet pas de représenter les phénomènes de dynamique discursive d'insertion et de réduction auxquels sont parfois soumis les termes complexes en situation textuelle (1997 : 209).

L'Homme et *al* classent le bruit produit par les logiciels qui utilisent cette technique en plusieurs types. Parmi ceux-ci, citons les trois principaux :

a) categorial ambiguity;
b) matching pattern but non terminological combination;
c) faulty delimitation of terminological units (v. 1996 : 306).

[19] Le bruit est un ensemble de formes fautives relevées par le logiciel, tandis que le silence est constitué de l'ensemble des formes présentes dans le texte qui n'ont pas été repérées par le logiciel.

Le premier type de bruit fait référence au fait qu'une forme lexicale peut appartenir à plusieurs catégories grammaticales. Le deuxième rejoint ce qu'explique Perron (v. *supra*, p. 67), c'est-à-dire que le patron typiquement terminologique récupère également des suites fortuites de mots. Finalement, le troisième type concerne le mauvais découpage exécuté par le système qui est incapable de délimiter les frontières de la chaîne terminologique dans la phrase.

À cet égard, Otman note l'existence d'une variation pour tous les patrons. Reprenons les réalisations possibles mentionnées par l'auteur :

> a) des unités terminologiques reconnues comme telles;
> b) des syntagmes plus ou moins lexicalisés mais qui n'ont pas le statut de terme car ils ne correspondent pas à des unités de signification (exemple*: gros ordinateur de bureau*);
> c) des groupes nominaux à contenu libre (ex: *la mémoire exceptionnelle de ce système*);
> d) des agencements accidentels ou inopinés dus à des collisions collocationnelles (ex: **aventure à but* extraite de "*système de conception assisté de récits d'aventures à but d'enseigner l'écriture*") (1992 : 66).

Il est donc important d'effectuer une analyse syntaxique plus large du contexte où ces unités s'insèrent.

La grande majorité des logiciels de découpage terminologique utilisent cette technique à l'une ou l'autre des étapes du fonctionnement. C'est le cas notamment de l'automate d'extraction de termes complexes *Filtact* conçu par une équipe de recherche du Ciral[20]. Après avoir fait indexer le texte par un analyseur textuel et avoir généré des segments complexes à l'aide d'un générateur de collocations, *Filtact* « va reconnaître toutes les formes complexes selon les valeurs grammaticales sélectionnées préalablement. En réalité, *Filtact* procède par la négative en écartant les patrons syntagmatiques qui ne sont pas possibles (ou peu probables) pour un TC[21] » (Auger et *al.*, 1996 : 48).

[20] Centre international de recherche en aménagement linguistique de l'Université Laval.
[21] TC signifie terme complexe.

Reprenons la troisième technique proposée par Perron, c'est-à-dire le découpage basé sur la description linguistique. Il s'agit en fait d'une approche qui s'appuie sur divers éléments résultant d'une description morphologique, syntaxique, lexicale et textuelle des unités. *Nomino*, le logiciel de l'Office de la langue française réalisé par l'équipe du Centre d'ATO[22], constitue un exemple d'application de cette technique. Selon David et Plante, *Nomino*

> [...] met en œuvre un certain nombre de principes qui relèvent essentiellement de la linguistique. Parmi ceux-ci, l'analyse syntaxique est posée comme un préalable nécessaire à tout traitement portant sur la langue naturelle, en particulier pour la reconnaissance des unités nominales polylexicales dont la structuration relève de la syntaxe. Les hypothèses élaborées permettent de reconnaître de (nouvelles) unités et de motiver le statut de ces dernières (1990 : 85).

Dans une perspective historique, de Schaetzen qualifie ce logiciel de système avancé d'aide au travail du terminographe. Selon l'auteure, « [...] le logiciel qui a fait date, ici, est incontestablement Termino[23], un détecteur de termes synaptiques [...] » (1994 : 61).

Présentement, la tendance, en ce qui concerne les nouveaux logiciels dédiés au découpage, va vers la combinaison de plusieurs approches.

En réalité, différentes techniques sont utilisées tout dépendant de l'approche adoptée par le logiciel. Habert et Jacquemin divisent les divers types de traitements selon les techniques employées. Leur typologie est classée en traitements :

a) structurels basés sur des grammaires formelles de complexité variable;
b) statistiques et numériques;
c) mixtes (v. 1993 : 19).

[22] Analyse de textes par ordinateur de l'Université du Québec à Montréal.
[23] Termino est le nom de l'ancienne version de Nomino.

Ces mêmes auteurs font une comparaison entre les approches structurelles et statistiques et mentionnent que

> Puisqu'elles n'opèrent pas à partir d'une base de connaissance structurée et stratifiée, les approches statistiques [...] associent aisément différents niveaux d'analyse. Au contraire, les approches structurelles dissocient les traitements à chaque niveau et requièrent des connaissances linguistiques les plus complètes possibles pour chaque étape (*id*. : 21).

Il est évident que chaque approche comporte des avantages et des inconvénients qui doivent être évalués en fonction du produit envisagé. Selon Habert et Jacquemin,

> La force de l'approche statistique tient dans la mise en évidence de phénomènes linguistiques difficilement accessibles aux approches structurelles parce qu'ils ressortissent à plusieurs niveaux linguistiques et qu'ils proviennent de données massives. Sa faiblesse vient de la difficulté à exploiter des résultats généralement hétérogènes. Au lieu de porter sur la collecte de données linguistiques, la réalisation d'un analyseur statistique de la langue repose sur le choix et la mise au point de fonctions mathématiques, sur leur paramétrage judicieux et sur l'exploitation des résultats de la collecte (*ibid*.).

La classification de Habert et Jacquemin est semblable à celle de Drouin. Selon ce dernier, les logiciels peuvent être caractérisés en trois types principaux, à savoir

a) les logiciels basés sur un modèle linguistique;
b) les logiciels basés sur un modèle statistique;
c) les logiciels basés sur un modèle hybride (v. 1997 : 46-54).

En ce qui concerne le premier type, les logiciels les plus simples font appel à la reconnaissance des catégories grammaticales, au moyen d'un dictionnaire et d'un analyseur morphologique, et aux règles contextuelles pour le choix d'une étiquette. Un analyseur morphologique, également appelé analyseur lexical ou lemmatiseur, « a pour but de traiter les mots d'une phrase pour en déterminer les catégories grammaticales puis en extraire les

formes lemmatisées par analyse des flexions » (Otman 1991 : 71). Toutefois, les logiciels qui utilisent uniquement cette technique ne réussissent pas à rendre compte d'un grand nombre d'ambiguïtés. Par ambiguïtés, Otman entend, par exemple, les formes qui portent les mêmes suffixes mais qui appartienent à des catégories grammaticales différentes (ex: *parent, équivalent*) ou des formes qui appartiennent à plus d'une catégorie comme c'est le cas notamment de *ferme* qui peut être un nom, un adjectif ou un verbe (v. *id.* : 69). Il faut également tenir compte des difficultés liées à la définition du mot, puisque cette dernière est essentielle pour l'établissement des algorithmes qui règlent le fonctionnement de l'extraction terminologique. En effet, « The exact definition of what comprises a word will influence the recognition of compounds, loosely bound prefixes and acronyms » (Lauriston, 1994 : 149).

Afin de résoudre ces ambiguïtés, il existe des logiciels plus élaborés qui utilisent un analyseur syntaxique qui « cherche [...] à générer des structures syntaxiques de phrases selon une grammaire donnée écrite sous la forme de règles de production » (Otman, 1991 : 71-72). L'Homme et *al* ajoutent que « The syntactic analysis can be simple (analysis of the local context in which the ambiguous form appears) or complex (complete parsing of the sentence) » (1996 : 297). Toutefois, ces logiciels peuvent, eux aussi, présenter des problèmes d'ambiguïté tel que la reconnaissance de termes et de non-termes. Pour les solutionner, les concepteurs de logiciels doivent s'appuyer sur des analyses sémantiques ou sur des techniques statistiques.

Même si l'approche sémantique est la plus acceptable, du point de vue linguistique, elle demeure encore très coûteuse et peu efficace quant aux résultats produits. Drouin constate que « À court terme, le couplage d'un module sémantique ne peut être une solution économiquement envisageable dans le cadre de la mise en place d'un système de reconnaissance des unités terminologiques complexes qui serait appelé à traiter des textes provenant de domaines très différents » (1997 : 46). Bien que rares, on note que certains efforts sont faits pour mettre à profit ce type d'approche.

L'un des logiciels fortement axés sur une approche linguistique est le logiciel *ZStation*[24]. Il s'agit en réalité d'un environnement informatique basé sur l'intelligence artificielle. Parmi les diverses fonctions offertes dans *ZStation*, il y a l'extraction de mots composés et de structures phraséologiques. *ZStation* a été conçu suivant le principe « qu'une application de génie linguistique requiert généralement des ressources linguistiques; si celles-ci ne sont pas disponibles, l'utilisateur doit pouvoir les créer lui-même » (Zinglé, 1996 : 80). Ces ressources linguistiques sont :

- les dictionnaires
- les ontologies
- les grammaires de flexion
- les grammaires pour l'analyse syntaxique
- les ressources pour la dérivation morphologique
- les ressources pour la composition morphologique
- les grammaires de spécification pour diverses activités (*ibid*.)

Selon cet auteur, pour ce qui est de « l'extraction des composés, deux méthodes sont actuellement utilisées dans le cadre de la ZStation, fondées respectivement sur une approche stochastique et sur un ensemble de spécifications linguistiques » (*id.* : 85). Le processus de spécifications linguistiques, plus particulièrement, « est fondé sur une grammaire élaborée par l'utilisateur dont l'objectif est de décrire les structures de composition au plan formel » (*ibid*.)

Le logiciel *Nomino*, cité plus haut (v. *supra,* p. 70), est un autre exemple d'application qui privilégie l'analyse de base linguistique. « Il est essentiel de préciser ici que les concepteurs de TERMINO[25] n'ont pas retenu la stratégie du recours aux patrons et modèles linguistiques qu'ils estiment non satisfaisante sur le plan linguistique. Ils s'appuient donc strictement sur une analyse syntaxique des textes pour repérer des unités linguistiques (les synapsies) potentiellement qualifiables d'unités terminologiques par un

[24] ZStation a été créé par le Laboratoire d'ingénierie linguistique et de linguistique appliquée de l'université de Nice.
[25] En majuscule dans le texte.

terminologue humain » (Otman, 1991 : 84). En effet, *Nomino* est un logiciel qui intègre la reconnaissance des unités complexes à l'analyse syntaxique.

En ce qui a trait à l'approche purement statistique, Daille mentionne qu' « Avec l'arrivée de techniques permettant la gestion et le traitement de grands corpus, les méthodes statistiques sont devenues très présentes dans le traitement du langage naturel. Ignorées ou mésestimées par les linguistes, celles-ci occupent néanmoins une place de plus en plus importante » (1994a : 8).

Selon L'Homme et *al*, « This approach is based on the fact that words that co-occur more frequently than can be expected [...] must carry significant information. In other words, frequently co-occurring lexical items cannot be accidental: a particular usage must be responsible for this frequency » (1996 : 298).

Kageura et Umino relèvent quatre types de fréquence d'usage :

1. A unit which appears frequently in a domain is likely to be a term of that domain.
2. A unit which appears only in one domain is likely to be a term of that domain.
3. A unit which appears relatively more frequently in a specific domain than in general is likely to be a term of that domain.
4. A unit whose occurrence is biased in some way to (a) domain(s) is likely to be a term (1996 : 280).

Cependant, mesurer la fréquence optimale demeure une tâche difficile puisqu'il s'agit d'un raisonnement essentiellement intuitif.

Dans le cas de l'analyse des séquences de mots, les mesures statistiques utilisées sont fondées sur le degré d'association entre les mots. Très employées pour l'extraction des patrons co-occurrents, ces techniques d'attraction entre les mots appliquent essentiellement deux types de calculs, à savoir : l'information mutuelle et le score d'association. Selon McEnery et Wilson, « The mutual information score between any given pair of words - or

indeed any pair of other items such as, for example, part-of-speech categories - compares the probability that the two items occur together as a joint event (i.e. because they belong together) with the probability that they occur individually and that their co-occurrences are simply a result of chance » (1996 : 71). Quant à la technique relative au score d'association, ces auteurs mentionnent qu'elle « [...] provides similar data to mutual information. For any given word (or other item) in a text, this test compares the actual frequency of all other words occurring within a specified size of context window (for example three words either side of the item) with their expected frequency of occurrence within that window if only chance were affecting the distribution » (*id* : 72).

En comparant l'approche linguistique à celle fondée sur la statistique, Drouin (v. 1997 : 47) présente les avantages de cette dernière, que nous résumons dans le tableau ci-dessous :

Tableau 2.3 — Comparaison entre les approches linguistique et statistique

approche linguistique	**approche statistique**
manipule des mots	manipule des chiffres rattachés aux mots
demande beaucoup d'appel à la récursivité	ne demande pas autant d'appel à la récursivité
moins rapide	plus rapide
nécessite beaucoup de ressources logicielles et d'injection de connaissances, p. ex: dictionnaires pour la lemmatisation, pré-traitement humain du corpus, règles grammaticales, etc.	nécessite moins de ressources logicielles et d'injection de connaissances
dépendance du système par rapport au domaine	n'a pas une dépendance par rapport au domaine

À notre avis, la statistique est un moyen valable mais qui ne doit pas être considéré comme étant la seule approche qui existe quand on traite de l'objet linguistique. De plus, même les études purement statistiques de la langue ont elles aussi besoin d'être interprétées. Les données numériques ne représentent pas grande-chose à elles seules, il faut les

expliquer. C'est ainsi que le lexicologue et statisticien Muller fait référence à son propre métier en ces termes :

> On se figure généralement qu'une statistique est une accumulation de chiffres offerts à l'administration du public. Or le comptage, qu'il soit long ou bref, n'est que le premier pas d'une opération statistique. Il est rare que par lui-même il fournisse une conclusion utilisable. [...] il faut le considérer comme une *observation*, qui n'a de prix que si elle peut ensuite être comparée à un *modèle*. (1979 : 215).

Par conséquent, l'analyse linguistique est toujours présente quelle que soit l'approche préconisée. Nous remarquons qu'une grande partie des logiciels destinés à l'extraction de termes est conçue selon un modèle hybride, autrement dit dans une perspective autant linguistique que statistique. La seule différence concerne la préférence donnée comme point de départ, c'est-à-dire où le logiciel commence par l'analyse linguistique et utilise les calculs statistiques comme méthode de filtrage et de correction ou bien le système adopte une démarche statistique au départ et ensuite utilise des filtres linguistiques. Bourigault, pour ne nommer que lui, appuie l'élaboration des logiciels du premier type :

> Les critères de fréquence ne sont, à notre avis, pas les meilleurs critères à prendre en compte *en premier lieu* pour le repérage des termes. [...] il est particulièrement important et efficace d'utiliser, après l'analyse syntaxique, certain calculs statistiques, correctement définis, sur la liste des candidats termes obtenus, pour effectuer un filtrage et un typage des candidats termes (1993 : 114-115).

Bourigault a adopté cette démarche pour la conception du logiciel d'extraction de termes *Lexter*. Ce logiciel a comme principe de base le découpage du texte par repérage de frontières. Ainsi,

> Plutôt que d'exploiter des connaissances "en positif" sur les structures grammaticales possibles des termes complexes (*nom de nom*, *nom adjectif*, etc.), il s'agit au contraire de s'appuyer sur des connaissances "en négatif" concernant les configurations grammaticales dont on sait qu'elles *ne* peuvent

> *pas* être des constituants de termes (*verbe*, *conjonction*, *pronom*, etc.) (*id* : 108).

D'ailleurs, comme nous l'avons déjà mentionné (v. *supra* p. 69), le logiciel *Filtact* utilise lui aussi la technique d'exploitation en négatif.

Habert et Jacquemin, dans un commentaire sur le système conçu par Bourigault, signalent que « Les marqueurs de frontières sont sûrs lorsqu'il s'agit de verbes conjugués, mais sujets à caution dans le cas de prépositions telles que *sur une* dans *logiciel de traitement de texte sur une station de travail*. C'est par un paramétrage plus ou moins détaillé des données linguistiques spécifiques au domaine étudié que l'on augmente la précision de l'extraction » (1993 : 33).

Le système conçu par Daille s'insère dans cette ligne de pensée qui unit des méthodes autant linguistiques que statistiques. Dans un premier temps, ce système procède à l'extraction des termes candidats à partir d'une typologie morphosyntaxique des termes. Dans un deuxième temps, la technique utilise des mesures statistiques qui calculent :

a) le nombre d'occurrences du couple;
b) un critère d'association : le coefficient de vraisemblance;
c) les mesures de distance (v. 1993, 1994a, 1994b, 1994c, 1995).

Le logiciel *Filtact* (v. *supra*, p. 69) emploie lui aussi l'approche mixte. Les diverses techniques appliquées dans ce logiciel sont, selon les auteurs, classées en

a) approche linguistique
   - catégorisations grammaticales et / ou syntaxiques des unités
   - applications de patrons de fouille

b) approche statistique
   - fréquence d'apparition des segments
   - calcul de l'attraction entre les lexies (score d'association)

c) approche linguistico-statistique
- conformité stricte à des règles structurelles syntaxiques
- fréquence d'apparition des segments
- coefficient de vraisemblance
- mesure de distance (Auger et *al*., 1996 : 50)

Le modèle hybride a également été employé dans la méthodologie proposée par Drouin. Cette technique est composée des étapes suivantes :

a) analyse statistique
- repérage des candidats-termes
- évaluation probabiliste du statut des candidats

b) analyse linguistique
- filtrage selon la morphologie des non-termes complexes
- analyse de l'autonomie des candidats
- analyse des candidats en contexte (1997 : 48).

Cette méthodologie est adoptée par le prototype *Notions* pour le dépistage des descripteurs complexes en langue de spécialité. Les concepteurs de ce prototype se sont fondés sur la ressemblance entre la terminologie et la gestion documentaire, c'est-à-dire sur un terme et un descripteur. Selon les auteurs « En gestion documentaire, notamment pour la constitution de bases de données bibliographiques, on élabore, pour représenter le contenu des documents, des listes de mots clés, ou de descripteurs. En terminologie, on cherche notamment à identifier l'ensemble des termes qui caractérisent un domaine donné, domaine souvent circonscrit par un corpus de textes » (Ladouceur et Drouin, 1997 : 207).

Habert et Jacquemin mentionnent que la plupart des approches utilisées pour l'extraction se restreignent aux associations binaires. Ils expliquent que « Pour les associations de plus de deux mots, l'acquisition requiert un travail complémentaire ou les "composés" acquis initialement jouent le rôle d'amorce. Dans la mesure où une forte proportion de "composés" de plus de deux mots sont construits par surcomposition à partir de "composés" élémentaires binaires [...], leur acquisition s'obtient par élargissement des "composés" binaires » (1993 : 34).

Quelle que soit l'approche employée par le système de dépouillement terminologique, celui-ci demeure toujours un outil dont les résultats doivent être épurés par le terminographe. Afin de bien cerner l'utilité réelle d'un système, il faut l'évaluer selon les besoins propres à chacun. L'Homme et *al* suggèrent une grille d'évaluation des logiciels d'extraction terminologique. Pour ce faire, ces auteures établissent des critères externes (paramètres correspondant aux besoins de l'utilisateur) et internes (performance du système). Nous reproduisons ci-dessous leur grille (L'Homme et *al*. 1996 : 295).

Tableau 2.4 — Grille d'évaluation de L'Homme et al

Type of criterion	Criterion	Options
Pre-evaluation[26]	A. Basic design	a. application b. language(s) processed c. type of system
	B. Strategies for automatic term extraction	a. linguistic approach b. statistical approach
	C. Units identified	a. noun phrases b. others
	D. Varied options	a. frequency of combinations b. combinations previously identified
	E. Lists	a. number of lists b. form of lists

En ce qui concerne les critères internes, L'Homme et *al* analysent la capacité du logiciel de relever seulement des unités terminologiques effectivement recherchées par l'utilisateur. Selon eux « The ideal system should be able to extract all complex terminological units contained in a given text and extract those units only, but no existing term-extraction facility succeeds in this task without producing a certain quantity of errors » (*id*. : 302). L'étude retient des critères classiques d'évaluation tels que le bruit et le silence.

[26] *Pre-evaluation* renvoie aux critères externes

Finalement, ces auteures font une remarque judicieuse en ce qui a trait à l'interprétation des résultats de leur étude. D'après eux, « [...] the performance is determined by all sorts of factors, such as the length of the text, the type of text (specialized field covered), the language processed by the system, etc. The figures can also be influenced by the evaluators' definition of "terminological unit" » (*id.* : 303).

En guise de conclusion à cette section, résumons quelques pistes suggérées par Habert et Jacquemin à propos de l'élaboration des méthodes d'extraction. Ces chercheurs mentionnent qu'il faut :

- préciser avec plus de finesse les modes de formation des unités complexes;
- mieux articuler les règles générales de construction et d'interprétation du syntagme nominal avec les particularités des « composés nominaux », ainsi qu'avec les contraintes propres à chacun d'entre eux;
- prendre en compte les transformations morphosyntaxiques qui altèrent la chaîne graphique mais qui restent assez proches de l'interprétation initiale pour ne pas être l'occurrence d'une nouvelle entrée lexicale;
- mieux articuler et associer les diverses approches;
- adjoindre à des mécanismes linguistiques réguliers de construction de la forme et du sens des dispositifs d'inférence et de raisonnement basés sur un savoir pragmatique qui viendront compléter les ébauches d'interprétation produites par le module linguistique, et qui s'appuieront également sur des indices textuels multiples, au premier rang desquels les indices statistiques (v. *id.* : 35-40).

2.3.2 Traitement et récupération terminologique

Comme nous venons de le voir, les différentes approches utilisées par les logiciels d'extraction de terminologie ont recours à une série de techniques de manipulation de données très riches pour les besoins du travail terminographique. Ces données traitées par des lemmatiseurs, des analyseurs syntaxiques, des analyseurs sémantiques, ou par d'autres techniques, sont d'une grande utilité pour l'exploitation de l'information linguistique contenue dans les banques de terminologie.

Une fois qu'on a réussi à exécuter l'extraction de données terminologiques, il faut traiter ces données tout en tenant compte des besoins de récupération de l'information terminologique. Les fiches produites par les logiciels d'extraction terminologiques doivent absolument contenir les éléments nécessaires pour la poursuite des travaux terminographiques. Selon de Schaetzen, « Rien d'étonnant [...] à ce que des fiches terminologiques soient de plus en plus souvent créées comme sous-produits de la structuration des réseaux sémantico-cognitifs, de recherches statistiques, etc. » (1997 : 87).

Dans une analyse relative aux données des fiches terminologiques traditionnelles, Zinglé constate que le manque d'associations entre les rubriques implique nécessairement une perte de résultats intéressants qu'on aurait pu avoir dans une structure associée. Il cite, à titre d'exemple, que « [...] la relation de synonymie est généralement indiquée de façon explicite, alors qu'elle peut être déduite de l'association au même concept de deux termes répondant aux même critères de normalisation » (1997 : 8). Cette réflexion a donné lieu à la création du logiciel d'acquisition et de traitement des données terminologiques ZTermino[27]. Cet outil profite de l'environnement ZStation, un produit du « génie linguistique fondé sur les méthodes de l'intelligence artificielle distribuée et orientée vers la modélisation de connaissances linguistiques en vue d'applications liées au traitement automatique des langues » (Zinglé, 1996 : 79).

Nous croyons d'ailleurs qu'au cours des diverses étapes exécutées par les logiciels d'extraction terminologique il y a sûrement une quantité considérable de données utiles pour le traitement de la terminologie par le terminographe qui demeurent en suspend. Par conséquent, nous considérons que les extracteurs terminologiques doivent envisager l'intégration et l'articulation des données manipulées au moment de l'analyse terminographique. Une simple liste de termes extraite sur une base lexicométrique de la

[27] ZTermino a été conçu par le Laboratoire d'ingénierie linguistique et de linguistique appliquée de l'Université de Nice.

cooccurrence des termes, par exemple, constitue une source d'information intéressante sur les termes rattachés à un réseau notionnel. À ce sujet, Otman précise que « L'analyse des mots associés consiste à calculer un indice d'association à partir de la co-occurrence de paires de mots dans un article. La conjonction de ces paires (rangées par ordre de similarité) permet la mise en évidence de "mots-carrefour" et de chaînes pour produire des réseaux » (1989 : 85-86).

Le logiciel *Pluritermes* créé par une équipe du Ciral constitue un exemple de système qui, en utilisant les résultats d'un logiciel d'extraction terminologique, est capable de repérer automatiquement des termes formellement semblables et également de générer des réseaux notionnels propres à un domaine (v. Ladouceur et *al* : 1997).

À notre avis, un système de traitement des données terminologiques doit donner accès à cinq niveaux d'analyse linguistique pour le terminologue, à savoir :

a) niveau morphologique (qui consiste à mettre en évidence les éléments de composition du terme hors contexte);
b) niveau syntaxique (qui considère les rapports entre les éléments, soit à l'intérieur des unités complexes soit dans des phrases, pour déterminer leurs positions et les représenter sous forme d'arbres syntagmatiques);
c) niveau sémantique (qui vise la représentation notionnelle interne véhiculée par les unités terminologiques);
d) niveau pragmatique (qui a pour but de rendre transparent les rôles extra-linguistiques des unités terminologiques dans le contexte discursif).
e) niveau sémantico-conceptuel (dont l'objectif est de mettre en évidence la structure terminologique d'un domaine à partir de traits liés à des concepts et notions extra-linguistiques).

Chaque niveau décrit ci-dessus a pour fonction d'éclaircir le niveau précédent. Ainsi, la syntaxe doit désambiguïser la morphologie; la sémantique enlève les ambiguïtés

de la syntaxe et la pragmatique celles de la sémantique. Au fond, cette façon de concevoir le traitement de la description linguistique est présent dans divers systèmes fondés sur des données textuelles. Tels sont, par exemple, les hypothèses qui ont guidé la méthode de construction des filtres linguistiques présentée par Naulleau et *al.* :

> a) [...] textual processing based on syntax can reveal semantic pattern and classes [...], and textual processing based on the combination of syntax and semantic may reveal conceptual information;
> b) [...] the search of terminological information with a computer cannot come from a conventional reading of the text, but may arise from specific investigation and organisation processes, such as syntactic normalization of phrases, normalisation of variants;
> c) [...] it is possible to distinguish them from common nominal phrases, if they are described as rich linguistic objects especially using semantic information (1996 : 808).

Les cinq niveaux du traitement des données terminologiques cités plus haut ont plusieurs fonctions, entre autres, l'aide à l'élaboration de définitions, la possibilité de recherche terminologique au moyen des structures notionnelles (recherche onomasiologique), l'élimination des redondances de la banque, l'élaboration de produits terminologiques structurés notionnellement et l'aide aux études linguistiques.

En effet, le traitement de données doit envisager non seulement la recherche du terme mais également l'environnement linguistique et fonctionnel de la langue. Dans cette perspective, Gouadec avait déjà suggéré un système fondé sur « comment 'fonctionne' le terme dans la langue et comment 'fonctionne' le concept ou l'objet désigné dans la pratique et dans l'expérience » (1987 : 132).

Dans la pratique du travail terminographique, le traitement des unités terminologiques se fonde sur des textes. L'automatisation de cette étape est fondamentale et les outils informatiques offerts sur le marché sont multiples. Les analyseurs de textes constituent des outils indispensables pour le traitement du texte aux fins terminographiques. En effet, ils font le repérage et le découpage des contextes qui seront ensuite soumis à un

système de gestion des contextes pour en effectuer l'analyse sémantique. Cette dernière étape est un outil important d'aide à la rédaction de définitions et à la gestion d'arbres de domaines (v. Auger et *al*., 1991 : 125). D'ailleurs, Auger et *al*. signalent que « Les modules de gestion de l'information et d'aide à la rédaction de l'information, pour certaines étapes du travail, comme par exemple l'aide à la rédaction des définitions et des notes, [...] [fournissent] des pistes intéressantes dans le contexte de l'automatisation de la chaîne de travail en permettant une maximisation de l'information extraite à diverses étapes du travail » (*id*. : 671-672).

Plus précisément, en ce qui concerne la sémantique conceptuelle, les notions représentées en structures arborescentes facilitent la définition des termes par corrélation entre notions aussi bien que le regroupement des termes par domaines de connaissance. Aujourd'hui, les banques de connaissances terminologiques sont le support le plus convenable pour ce type de structuration terminologique. C'est d'ailleurs de cela que nous traiterons dans la prochaine section.

2.3.3 Base de connaissances terminologiques

À la fin des années 60, grâce à l'avènement de l'informatique, les premières banques de terminologies ont vu le jour. La naissance des banques comme Eurodicautom de l'Union européenne, la BTQ de l'Office de la langue française du Québec et Termium du Gouvernement fédéral canadien, ne sont que quelques exemples. À cette époque les banques de terminologie étaient supportées par un seul gros ordinateur et manipulaient une grande masse de données. L'informatique était donc perçue comme une solution au problème de la quantité d'information. La qualité de l'information saisie et les moyens de la récupérer ne faisaient pas encore partie des préoccupations majeures. En réalité, les progrès de l'informatique n'étaient pas suffisants pour la création de techniques avancées de traitement et d'accès à l'information.

Au cours des années 80, on a vécu une transformation notable du type de traitement du contenu des banques. D'une conception centralisatrice des grandes masses de données enregistrées sur un seul ordinateur, on est passé à la conception de petites banques, plus spécialisées. Cette nouvelle génération a également contribué au changement de leur fonction. Au début considérées comme instruments de diffusion et de consultation, les banques sont devenues de véritables instruments de production de terminologies.

Toutefois, le progrès informatique de notre fin de siècle a apporté plus qu'une simple addition de fonctions, nous avons aujourd'hui la possibilité d'automatiser le travail terminographique à partir d'une approche onomasiologique. Enfin, les terminographes peuvent adopter une démarche structurée multirelationnelle de l'objet de la terminologie. Ce sont d'ailleurs les recherches menées en intelligence artificielle qui ont rendu possible cette nouvelle conception des banques de terminologies appelé *bases de connaissances terminologiques*.

Nous croyons que dès l'étape du dépouillement terminologique des mesures pour le traitement du sens dans le but de constituer une base de connaissances terminologiques devront être envisagées. Il est évident que l'exploitation des systèmes experts créés pour les banques de terminologie demande une révision globale de la structuration de ces types de banques depuis l'enregistrement de l'information terminologique jusqu'à la diffusion et à la récupération. Cependant, sur le plan de la méthodologie privilégiée par les terminologues travaillant avec une banque de connaissances, elle reste presque la même. En réalité, comme l'affirment Meyer et *al*., « Because concept management has always been a crucial component of terminologist work: they are keenly aware of the importance of certain depth of understanding of the concepts of the subject field, which implies the inter-relations between these concepts, for producing high-quality terms record » (1991 : 241).

Une banque de connaissances terminologiques comporte tous les éléments habituellement emmagasinés dans une banque terminologique traditionnelle enrichie d'informations liées à la structure conceptuelle d'un domaine. Plusieurs de ces banques

appliquent les modèles relationnels de réseaux sémantiques. Ces modèles partent du principe que la connaissance n'est pas formée d'éléments isolés mais, au contraire, que son organisation dans la mémoire humaine est formée d'une association d'éléments qui ont des rapports entre eux. C'est justement cet aspect essentiel de la relation qui est présent dans la représentation en réseaux sémantiques. À ce propos, Badia explique que

> Les objets mathématiques utilisés pour mettre en œuvre la connaissance sont les graphes, qui portent des adresses et des étiquettes, dont les nœuds représentent les concepts ou entités qui font l'objet de la connaissance représenté. De leur côté, les arcs étiquetés qui unissent les nœuds indiquent les relations entre ces concepts et ces entités. Les relations "is-a" et "part-of" utilisées dans les réseaux sémantiques se distinguent. Nous parlons plutôt de « frames » lorsque les nœuds qui composent les graphes ne sont plus de simples étiquettes, mais des structures de données complexes[28] (v. 1996 : 73).

La méthode des réseaux sémantiques, largement étudiée par Sowa (1984), est de plus en plus appliquée non seulement aux banques de terminologie, mais également à l'analyse de textes scientifiques. Rivalland, dans un article sur l'application de graphes conceptuels au discours scientifique, mentionne quelques avantages de l'utilisation de réseaux sémantiques. L'un de ces avantages c'est que ce modèle, par opposition aux dictionnaires conventionnels, rend la représentation de la définition dans une perspective relationnelle et dynamique possible. L'auteur ajoute que « Dans cette mesure les graphes sont plus près de la conception que l'on a aujourd'hui de la signification (notamment grâce aux apports de la pragmatique), selon laquelle un mot n'a pas en même temps *n* définitions, mais signifie ceci ou cela, en fonction des multiples liens qu'il noue avec le co-texte et contexte situationnel » (1993 : 66).

Otman présente quatre grands projets fondés sur les réseaux sémantiques : *Cogniterm*, *Matra Marconi Space*, le projet de traitement terminologique du Trésor de la

[28] Traduction personnelle du texte original rédigé en espagnol.

langue française (TLF) et les travaux de recherche du groupe Terminologie et intelligence artificielle (IA) (v. 1997 : 251-255).

Cogniterm est un projet pour la mise au point d'un hybride entre la banque de termes conventionnelle et la base de connaissances. Il a été conçu par le Laboratoire de l'intelligence artificielle de l'Université d'Ottawa (v. Meyer et *al*. 1991 : 232 et 233). Cet outil exploite l'environnement *CODE* (*Conceptually Oriented Descriptive Environment*) et suit une méthode divisée en trois étapes :

a) Établissement d'une esquisse d'un plan conceptuel. Cette étape consiste à l'identification des principaux sous-domaines et des limites entre domaines en utilisant les possibilités fournies par le graphe *CODE*.
b) Établissement d'une grille des caractéristiques pour chaque sous-domaine sélectionné. Cette étape a pour but de détailler la compréhension des domaines en partant d'une analyse approfondie du domaine plus fondamental jusqu'aux sous-domaine.
c) Lecture optique de textes et analyse conceptuelle. À l'aide du système *CODE*, la documentation sur les sous-domaines est lue par un lecteur optique. À chaque concept trouvé, le système crée un descripteur conceptuel, lequel est intégré dans une hiérarchie du type genre et espèce (v. *id*. : 241-242).

Le projet *Matra Marconi Space*, conçu par la société Matra Marconi Space et l'équipe de Recherche en Syntaxe et en Sémantique de l'Université de Toulouse-Le Mirail, est fondé « à la fois sur des méthodes de sémantique des concordances et de celles de l'acquisition des connaissances de l'intelligence artificielle » (Condamines 1994 : 48). L'analyse linguistique est réalisée au moyen d'outils informatiques comme les extracteurs de termes candidats, les outils d'analyse de contenu, les concordanciers et les gestionnaires de terminologies. L'objectif est d'analyser le fonctionnement des termes, c'est-à-dire d'analyser les concepts par rapport aux contextes où ils apparaissent. Cette analyse est faite

tant sur le plan syntagmatique que sur le plan paradigmatique. Le résultat de l'analyse est une ébauche de réseau conceptuel soumis à l'approbation des experts.

Quant au projet de traitement terminologique du TLF, il consiste à introduire les termes repérés et les relations qui existent entre eux « dans un modèle formel de représentation des connaissances. L'intérêt d'un tel modèle dans le cadre de ce projet est de donner aux utilisateurs un moyen de "raisonner" sur le contenu de la base en associant aux relations des propriétés logiques, et d'interroger la base sur l'interdépendance des termes dans le dictionnaire » (Otman 1997 : 254).

Finalement, les travaux de recherche du groupe Terminologie et IA constituent des réflexions quant à la réalisation d'une banque de connaissances terminologiques dans le domaine médical. Otman mentionne que les activités sont fondées principalement sur

a) la distinction entre les notions de base de connaissances terminologiques et systèmes à base de connaissances pour que la première puisse être exploitée par le second;
b) l'interface entre base de connaissances terminologiques et traitement de la langue naturelle (TLN) étant donné que « les systèmes de TLN ont besoin de "connaissances terminologiques" au moment de l'analyse syntaxique (reconnaissance des UT) et de l'analyse sémantique (connaissances sémantiques véhiculées par les termes) pour interpréter les phrases et les textes qui leur sont soumis » (*ibid*.);
c) la création d'un formalisme de représentation des connaissances (*id*. : 255).

Dans cette même veine d'outils qui privilégient l'approche fondée sur la représentation des connaissances, citons le projet terminologique multilingue du logiciel ZTermino (v. *supra*, p. 81). Ce logiciel « [...] permet en particulier de traiter chaque langue séparément et d'opérer la mise en regard des termes via les concepts non seulement dans une langue particulière mais également de langue à langue. Cette approche a également des incidences intéressantes du point de vue de la traduction à base de connaissance et du traitement documentaire multilingue » (Zinglé 1997 : 9).

En guise de conclusion à cette section, nous dirons que, d'une part, l'expertise acquise par la constitution des bases de connaissances terminologiques aidera à améliorer les performances des logiciels d'extraction de terminologie puisqu'elle apportera une connaissance linguistique (analyse syntaxique et sémantique) essentielle pour le fonctionnement de ces types d'outils terminotiques. D'autre part, la conception de logiciels d'extraction doit prévoir l'alimentation de ces bases en termes de données relatives aux connaissances descriptives de l'objet linguistique.

2.4 Problématique de la recherche

Grâce aux progrès de l'informatique, les banques de données terminologiques sont apparues comme une solution pour l'enregistrement et la diffusion de la production terminologique d'un pays. Le traitement automatisé des données terminologiques permet la mise à jour rapide et efficace de la terminologie créée au sein des langues de spécialité.

Conscients de l'importance de ces outils terminologiques, l'Institut brésilien d'information en science et technologie - IBICT et l'Université de Brasília (Brésil) ont mis sur pied, en 1994, le *Projet intégré pour l'implantation et la diffusion de la terminologie scientifique et technique au Brésil* (v. Faulstich et Café, 1994). Ce projet prévoit la création de la Banque de données terminologiques du Brésil (Brasilterm).

L'implantation de cette banque permettra d'atteindre les buts suivants :

a) diffuser la terminologie en langue portugaise, de façon à consolider son usage comme langue de communication scientifique et technique;
b) aider à perfectionner la traduction assistée par ordinateur;

c) servir d'instrument de base pour la création de la banque de données terminologiques du Mercosul[29].

Cependant, pour arriver à définir la structure de la banque de données (types de champs, types des données à recueillir, types de services et produits à offrir, etc.), nous devons, d'une part, connaître les caractéristiques fondamentales des unités terminologiques qui forment les textes de langues de spécialité écrits en portugais du Brésil et, d'autre part, définir une manière d'exploiter ces caractéristiques au moyen d'une chaîne de traitement terminographique automatisée conçue spécifiquement à cet effet.

Selon Auger, la chaîne terminographique automatisée est composée de quatre grandes catégories de tâches :

a) la préparation du travail terminologique;
b) le travail terminographique proprement dit;
c) le stockage et le traitement des données recueillies;
d) l'édition et la diffusion d'un produit final (1994 : 50).

En ce qui concerne la deuxième tâche, l'auteur mentionne qu'elle a un rapport avec « l'accomplissement de tâches proprement lexicographiques ou terminographiques liées au traitement d'écrits généraux et spécialisés » (*ibid.*). En effet, elle consiste à traiter un corpus de façon automatique et à faire l'extraction des données terminologiques.

Toutefois, pour arriver à la conception d'un logiciel d'extraction terminologique, il est essentiel de connaître les processus de formations des unités terminologiques ainsi que leur délimitation à l'intérieur des textes spécialisés. En d'autres mots, la maîtrise des comportements linguistiques de la structure interne des termes complexes et la

[29] Comme l'*Union européenne* et l'*Aléna*, le *Mercosul* (Marché commun du sud) est un accord commercial entre des pays voisins. Il a été créé en 1991 et regroupe les pays suivants : le Brésil, l'Argentine, le Paraguay et l'Uruguay.

connaissance des adaptations de la terminologie vis-à-vis du contexte discursif constituent des priorités pour la mise sur pied d'un logiciel de cette nature.

Bien qu'elle soit aujourd'hui nécessaire et prioritaire pour la création de la banque du Brésil (Brasilterm), une recherche de cette nature n'a jamais été entreprise. Concrètement, nous pensons que la description exhaustive des UTC existantes dans les langues de spécialité en portugais du Brésil et l'analyse approfondie de cette description nous permettront d'atteindre notre objectif. Par son caractère inédit, cette recherche contribuera à l'avancement de la théorie de la terminologie

2.5 Objectif de la recherche

L'objectif général de cette recherche est de proposer des critères minimaux pour la conception du logiciel d'extraction de termes qui servira à l'alimentation de la banque du Brésil.

Pour atteindre cet objectif, nous décrirons et analyserons, à la lumière de la grammaire fonctionnelle, les caractéristiques morphosyntaxiques et syntactico-sémantiques de la terminologie de la biotechnologie des végétaux en langue portugaise. Plus particulièrement, étant donné la haute fréquence des UTC relevées en contexte, nous allons restreindre l'objet de notre recherche aux UTC de base nominale.

2.6 Conclusion

Au cours de ce chapitre, nous avons fait un survol de deux principaux aspects de l'état de la question. Premièrement, nous avons fait le point sur plusieurs recherches menées sur la formation des UTC et sur les questions liées au figement de ces unités. De tout ce qui a été dit sur ce sujet, nous pouvons conclure que l'analyse des UTC n'a pas de réponse unique, c'est-à-dire une solution morphologique, syntaxique ou encore seulement

sémantique. C’est pour cette raison que nous avons choisi une analyse de base fonctionnaliste qui prend en considération tous ces éléments dans leur ensemble. Dans notre conception, les UTC sont des unités formées d’une base suivie de prédications (nucléaires ou étendues), dont les composants véhiculent des fonctions morphologiques, syntaxiques, sémantiques et pragmatiques définies selon les rapports établis entre eux. Pour nous, une UTC obéit à un processus d’incidence des prédications successives et non pas seulement à une structure binaire comme l’ont constaté plusieurs auteurs, notamment Guilbert (v. supra, p. 25 et suiv.).

Deuxièmement, nous avons mentionné plusieurs approches pour l’extraction terminologique, ce qui nous a mieux fait voir l’extrême complexité du sujet. D’ailleurs, si l’analyse linguistique de l’objet terminologique est difficile à réaliser, son automatisation est encore plus complexe du fait que, dans ce cas, il faut réussir à faire « penser » une machine.

Ainsi, nous avons présenté un aperçu des logiciels qui allait des logiciels les plus simples, fondés sur des analyses purement syntaxiques, jusqu’aux logiciels avancés conçus dans le domaine de l’intelligence artificielle. Ces derniers, qui n'en sont encore qu'à leurs débuts, commencent à avoir de plus en plus leur place dans le domaine de la terminotique.

Nous avons ensuite fait part de la problématique qui nous a incitée à poursuivre cette recherche. Comme nous l’avons mentionné, le projet de la banque de données terminologiques du Brésil a été l'élément déclencheur de cette démarche.

Compte tenu de l’ampleur d'une recherche de cette nature, nous avons établi des objectifs précis en fonction de nos besoins. Ainsi, partant du fait que l’extraction est l'une des premières étapes de la chaîne terminographique, notre principal objectif a été de proposer des critères pour l’établissement d’un logiciel de repérage. Dans ce contexte, l’analyse linguistique de la terminologie est fondamentale et elle constitue notre but spécifique. Pour y parvenir, nous avons choisi d'adopter la méthode fonctionnaliste d'analyse de la langue.

3. Constitution et informatisation du corpus

3.1 Introduction

Il sera question, dans le présent chapitre, de l'élaboration du corpus de textes qui seront étudiés et des divers traitements informatisés effectués. Nous présenterons d'abord la typologie des textes choisis ainsi que le chemin parcouru pour le repérage de ces textes. Par la suite, nous décrirons les étapes de l'informatisation du corpus et de l'analyse des données tout en justifiant le choix des logiciels utilisés.

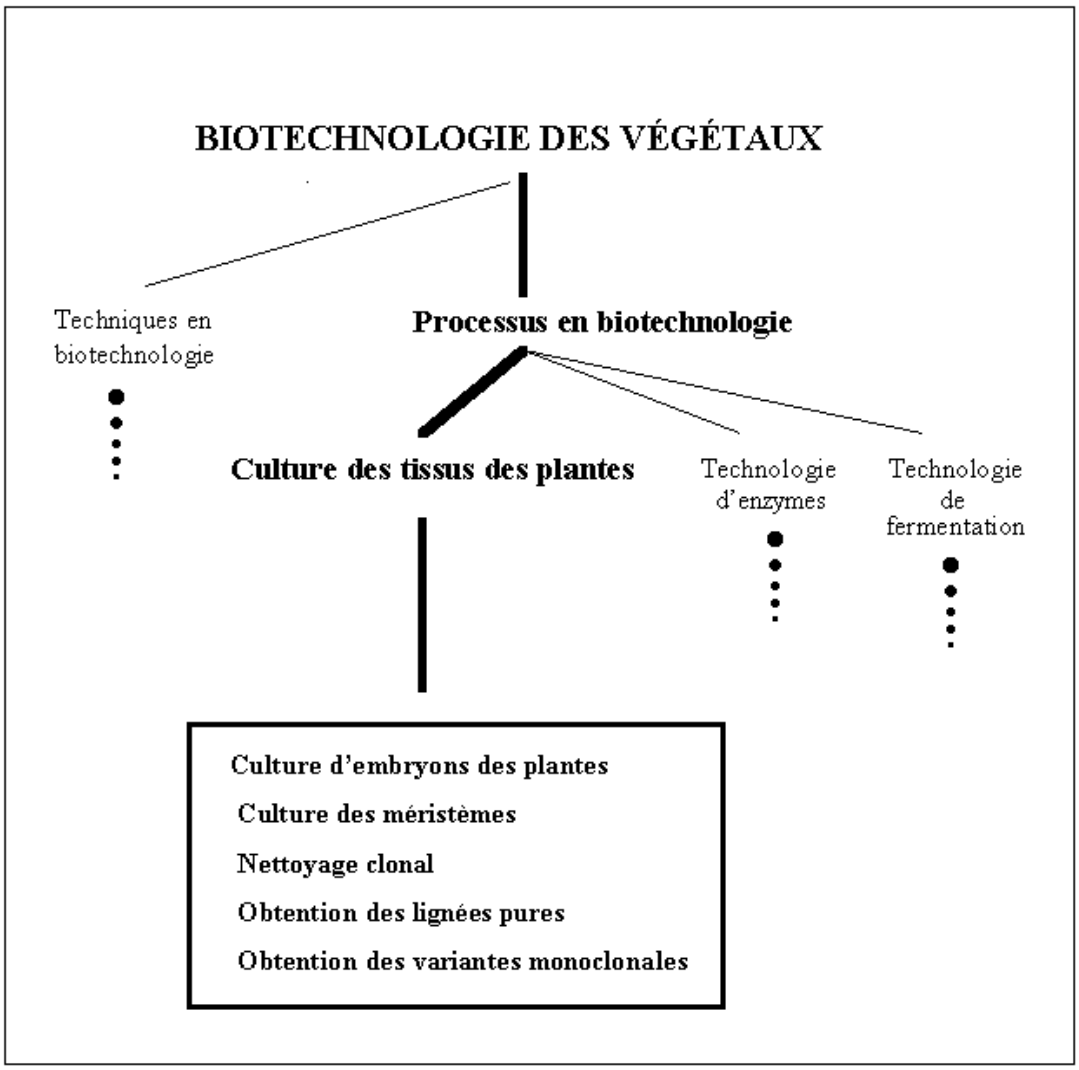
3.2 Choix du domaine

Nous avons choisi, comme domaine d'analyse pour notre étude, la biotechnologie et, plus spécifiquement, la biotechnologie des végétaux. Ce choix répond à des raisons extra-linguistiques et linguistiques.

Pour des raisons économiques, la biotechnologie est présentement considérée comme un domaine prioritaire tant par le gouvernement brésilien que par les pays du *Mercosul*. Étant donné que notre recherche vise à contribuer, dans un premier temps, à la conception de la Banque de données terminologiques du Brésil (Brasilterm) et, dans un second temps, à celle du *Mercosul*, nous avons décidé d'arrêter notre choix sur ce domaine de grand intérêt pour les gouvernements des pays du *Mercosul*.

Puisqu'il était impossible de traiter toute la terminologie relative à la biotechnologie, il a été nécessaire de procéder à un découpage thématique du domaine. Afin de rassembler une nomenclature terminologique cohérente, nous avons décidé de nous restreindre à la biotechnologie des végétaux. Dans le domaine de la biotechnologie des végétaux, nous avons retenu les processus relatifs à la biotechnologie des cultures des tissus des plantes ainsi que les diverses subdivisions de ces processus. La figure suivante présente le classement hiérarchique du domaine et des sous-domaines choisis pour la recherche.

Figure 3.1 — Classement hiérarchique du domaine de la biotechnologie des végétaux



Il s'agit donc d'un découpage thématique du domaine dont le but est d'aider à systématiser l'analyse des données. Sur le plan sémantico-lexical, cette manière de procéder nous assure que l'échantillonnage est représentatif du domaine, c'est-à-dire que la collecte des UTC reflète vraiment les sous-domaines choisis.

Du point de vue linguistique, la langue de spécialité de la biotechnologie des végétaux comporte une terminologie en expansion. Étant donné que de nouveaux éléments, produits et activités découlent de recherches et de découvertes constantes faites en biotechnologie, la terminologie de ce domaine doit continuellement s'adapter aux besoins de communication entre les spécialistes. Nous avons d'ailleurs remarqué que la langue portugaise s'inspire largement de la terminologie anglaise pour la création d'une bonne partie de ses unités terminologiques. Ce phénomène, appelé emprunt externe (v. Boutin-Quesnel et *al.*, 1985 : 26), est une conséquence du transfert technologique du pays producteur du domaine. Autrement dit, la terminologie étrangère accompagne la technologie importée et doit, à ce moment, être adaptée pour s'intégrer au système linguistique de la langue d'arrivée.

La terminologie du domaine de la biotechnologie des végétaux est également caractérisée par le phénomène de l'emprunt interne. Ce second type d'emprunt est en fait un transfert de termes d'un domaine à l'autre (v. *ibid*.). Dans le cas de la biotechnologie des végétaux, l'emprunt se fait principalement aux domaines de la biologie, de la botanique, de la chimie et de la génétique.

Qu'il s'agisse d'un emprunt interne ou d'un emprunt externe, le fait que le terme emprunté s'installe dans le système linguistique de la langue d'arrivée lui enlève le statut de terme étranger. À ce sujet Guilbert mentionne qu' « [...] un terme d'origine étrangère cesse d'être néologique à partir du moment où il est entré dans le système linguistique de la langue d'accueil, c'est-à-dire quand, précisément, il cesse d'être perçu comme terme

étranger » (1975 : 95). Pour ce qui est des critères d'intégration de ces termes dans la langue d'arrivée, Guilbert va plus loin et précise que,

> [...] [l'] installation doit être jugée du point de vue des processus linguistiques qui conduisent à cette intégration plutôt que décidée sur le seul critère de la mention dans les dictionnaires de langue; ceux-ci, en effet, peuvent refléter une doctrine d'accueil ou de refus à l'égard des termes étrangers qui n'est pas nécessairement fondée sur des critères objectifs d'intégration de caractère linguistique (*id.* : 95 et 96).

En effet, l'établissement de critères de reconnaissance des termes néologiques n'est pas une tâche facile, principalement dans le cas de la terminologie de la biotechnologie en langue portugaise. L'absence de dictionnaires portugais récents spécifiques à ce domaine empêche souvent la vérification du figement terminologique dans le système linguistique portugais. Vu que le vocabulaire spécialisé en langue portugaise est truffé d'emprunts à l'anglais, les spécialistes de langue portugaise publient la majorité de leurs textes dans cette langue étrangère. Ce fait freine la création de dictionnaires spécialisés en portugais, puisqu'à première vue les besoins ne semblent pas suffisants pour justifier l'établissement d'une terminologie propre à la langue portugaise. Pourtant, il en va tout autrement dans la réalité.

En outre, il était impossible d'utiliser le critère de fréquence dans notre recherche puisque nous travaillons sur un nombre restreint de textes. En effet, nous avons adopté une perspective qualitative plutôt que quantitative. Donc, pour ce qui est de la détermination de la valeur réelle du figement des UTC recueillies, nous nous sommes appuyée principalement sur l'analyse fonctionnaliste des rapports syntaxiques, sémantiques et pragmatiques qui existent entre les différents éléments des UTC. Ainsi, nous adoptons la position de Guilbert (v. *ibid.*), qui considère que l'analyse du processus linguistique est une méthode fiable pour la vérification du processus d'installation des termes dans le système linguistique d'une langue.

Compte tenu de ces éléments, nous croyons que la flexibilité et la créativité des mécanismes linguistiques utilisés pour la biotechnologie des végétaux en langue portugaise font de ce domaine un domaine riche en variables linguistiques et terminologiques indispensables à notre recherche.

Maintenant que nous avons précisé la structure thématique du domaine, nous allons expliquer, dans la prochaine section, notre choix quant aux types de documents utilisés pour la collecte des données.

3.3 Typologie des textes

L'UTC est l'unité par excellence des notions propres aux langues de spécialité. C'est par le biais de ces unités que le spécialiste véhicule ses connaissances sur les plans oral et écrit. Elles ont pour fonction de préciser le contenu de la communication. En outre, elles fonctionnent comme des modes d'expression propres au domaine auquel elles appartiennent, ce qui fait qu'elles comportent un ensemble de caractéristiques morphologiques, syntaxiques et sémantiques qui les distinguent des mots composés de la langue commune.

Dans le milieu scientifique et technique, on trouve souvent des UTC dans les divers supports écrits. Selon Kocourek, « Le processus de la connaissance approfondie est concrétisé, saisi, rendu accessible sous la forme des textes spécialisés. Aussi les textes spécialisés sont-ils l'expression concrète de la connaissance approfondie que les spécialistes peuvent communiquer au public restreint de leurs pairs » (1991 : 33).

Même si on reconnaît l'importance capitale de la communication orale, le texte écrit est encore aujourd'hui le moyen privilégié pour communiquer les connaissances entre les spécialistes d'un domaine. En effet, chacun de ces supports requiert un type d'étude qui est de nature différente. À ce propos Kocourek mentionne que

> Le parlé et l'écrit sont deux représentations différentes, jouissant, dans la communication, d'une autonomie relative. Une fois disponible, l'écrit peut, lui aussi, servir de porteur de sens, indépendamment et sans intermédiaire du parlé. Ceci rétablit l'équilibre entre l'écrit et le parlé, permettant de les voir comme à la fois étroitement liés et relativement indépendants (*id*. : 27).

En réalité, les textes écrits sont plus rigides que la communication orale en ce qui a trait au langage. À ce sujet, Gaudin affirme que « [...] les articles scientifiques présentent une vision "canonique" de la vie de la recherche, du chercheur et du laboratoire. Et il faut bien voir que cela a d'importantes conséquences terminologiques : on ne voit jamais ni les concepts ni les signifiants à leur stade premier - tout est déjà restructuré » (1993 : 154).

Toutefois, il faut mentionner également que cette remarque n'atténue pas l'importance des textes écrits, car ils demeurent des outils très valables de transmission de la pensée innovatrice et fertile dans les langues de spécialité.

En effet, les recherches menées en terminologie ont recours aux deux types de situations de communication, c'est-à-dire l'oral et l'écrit. Dans une étude sur l'implantation terminologique, Rouges-Martinez et Fossat présentent une classification intéressante des communications orales et écrites utilisées comme source de collecte de données pour leur recherche. Ils les ont définies « [...] autour de deux pôles : des situations d'échanges formels, qui requièrent plus de rigueur dans l'expression, et des situations d'échanges informels où une certaine spontanéité peut reprendre ses droits » (1994 : 124).

Tout en considérant cette classification, nous nous sommes uniquement intéressée, dans notre étude, aux UTC relevées dans les textes écrits. En ce qui concerne le niveau de langue, nous avons privilégié les textes destinés aux spécialistes, aux professeurs et aux étudiants étant donné que c'est le public-cible de la Banque de données terminologiques du Brésil. Ainsi, parmi la masse de documents écrits utilisés et produits par ce public, nous en avons retenu trois types, à savoir :

- les livres (manuels);
- les articles de revues scientifiques;
- les rapports de recherche.

La période de publication des documents retenus s'étend de 1991 à 1996 afin d'assurer une cohérence terminologique du domaine choisi.

Après avoir défini les types de textes nécessaires pour notre recherche, nous avons contacté le *Centre national de la recherche en ressources génétiques et de la biotechnologie* (*Cenargen*[30]) qui relève de la *Société brésilienne de recherche en agronomie et sur les bovidés* (*Embrapa*[31]) afin d'obtenir de la documentation spécialisée. Il faut préciser que cette société est la plus importante au Brésil dans le domaine de la biotechnologie des végétaux. En effet, le *Cenargen* possède et produit la majorité de la documentation brésilienne sur ce sujet.

Nous avons donc demandé à l'*Embrapa* si nous pouvions avoir accès à sa documentation pour constituer notre corpus. Cette institution nous a fait parvenir sa base de données bibliographiques en disque optique compact (v. Embrapa, 1996). Cette base contient, entre autres, toute la production écrite en diverses langues par des spécialistes brésiliens et publiée par l'*Embrapa* et d'autres organismes. À partir de cette base de données, nous avons sélectionné six documents selon les critères suivants :

a) type de document;
b) date de publication;
c) langue;
d) sujet de recherche.

[30] *Cenargen — Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia.*
[31] *Embrapa — Empresa Brasileira de Pesquisas Agropecuária.*

Le corpus[32] est constitué d'ouvrages de types variés :

a) deux mémoires de maîtrise;
b) un ouvrage collectif composé de 18 articles de divers auteurs;
c) trois articles du type rapport de recherche tirés de revues scientifiques.

Ces documents sont des textes écrits originellement en portugais. Toutefois, puisque la terminologie relative à la biotechnologie des végétaux en langue portugaise contient un nombre remarquable de traductions, nous avons trouvé intéressant d'ajouter un document traduit de l'anglais vers le portugais. Il s'agit de la traduction d'un livre bien connu dans le domaine. Il faut signaler qu'avant d'utiliser des données extraites de la traduction, nous avons minutieusement comparé la traduction à l'original pour vérifier sa qualité.

Après avoir défini les documents pour la collecte des données, nous sommes passés à l'étape de l'informatisation de cette documentation dans le but de rendre la recherche la plus efficace possible.

3.4 Informatisation du corpus

L'informatisation du corpus nécessite trois étapes :

a) la numérisation des textes et la reconnaissance des caractères;
b) la création de la base textuelle indexée;
c) le repérage des UTC.

Premièrement, nous avons numérisé, à l'aide d'un lecteur optique, tous les textes qui étaient en format papier. Nous avons ensuite utilisé le logiciel *Omnipage Direct* pour

[32] La référence complète de chaque document figure dans la bibliographie sous la rubrique *Ouvrages constituant le corpus*.

reconnaître les caractères des textes numérisés. Une fois l'étape de la reconnaissance automatisée terminée, il a fallu relire et corriger chacun des textes parce que le logiciel ne reconnaît pas parfaitement tous les caractères.

Deuxièmement, nous avons inséré les documents en format Ascii dans le logiciel *Makebase*[33] (version 2.1.3 - 1994), un des programmes du progiciel *Tact*[34], qui génère une base indexée de données textuelles. C'est à partir de cette base que nous avons pu procéder à la troisième étape qui consiste à repérer les UTC dans les textes.

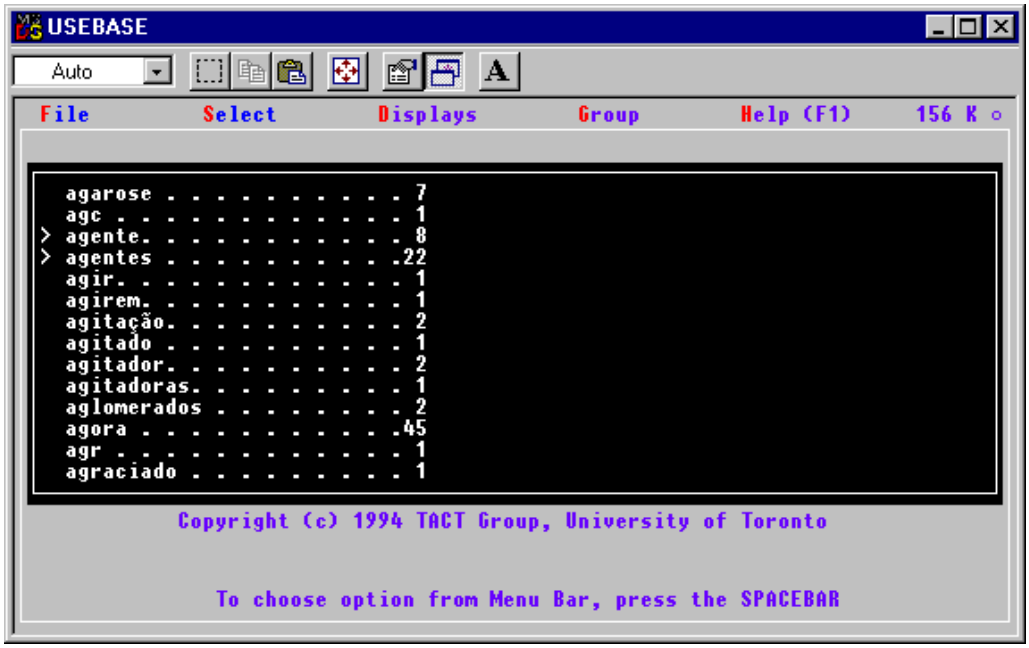
Cette troisième étape a été réalisée en mode semi-automatique grâce au logiciel *Usebase* (version 2.1.3 - 1994), un autre programme du progiciel *Tact*. Nous avons tiré avantage de ce logiciel pour repérer les UTC en contexte. En effet, *Usebase* produit des listes de termes simples et de termes complexes accompagnés de leur fréquence d'apparition dans les textes, ce qui permet de se reporter au contexte précis où le terme figure. Les étapes nécessaires pour obtenir ce résultat sont :

a) la sélection des termes simples;
b) le repérage de termes simples faisant partie d'une UTC;
c) la régénération de la base textuelle indexée;
d) la constitution des UTC suivies de leurs contextes.

Tout d'abord, nous avons sélectionné les termes qui pourraient faire partie d'une UTC soit comme base, argument ou satellite en utilisant la liste de termes simples fournie par *Usebase*. Voici un exemple d'écran produit par *Usebase* pour la sélection des termes simples. Les termes sélectionnés sont marqués par le symbole > .

[33] Le Tact Group de l'Université de Toronto a conçu le logiciel *Makebase* ainsi que le logiciel *Usebase*.
[34] Tact (Textual Analysis Computing Tools).

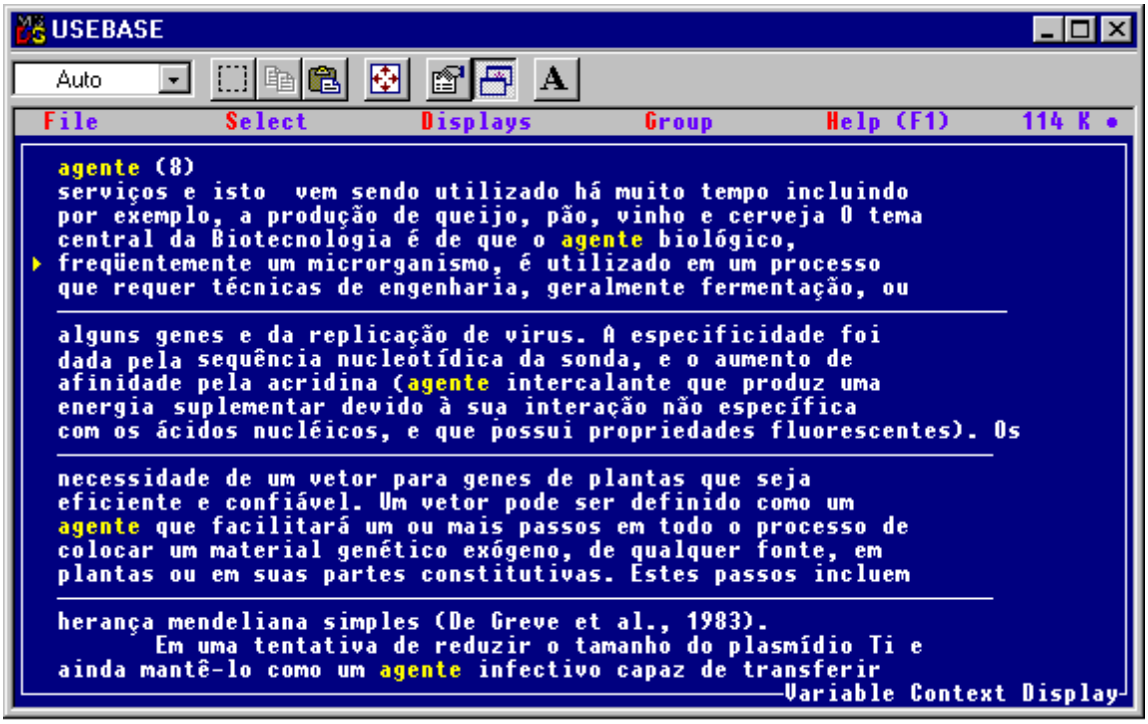
Figure 3.2 — Sélection des termes simples



À partir de la liste complète de mots simples produite par *Usebase*, nous avons retenu ceux qui constituaient des termes, par exemple *agente* et *agentes*, et nous avons vérifié, dans les contextes produits automatiquement par le logiciel, s'ils formaient une UTC avec les autres éléments environnants.

Nous présentons ci-dessous une partie de la liste de termes simples suivis de leurs contextes. Lorsqu'on examine la figure, on se rend compte que certains mots de la liste forment des UTC avec les mots avoisinants. Ainsi, dans le premier contexte, le terme *agente* est suivi de *biológico* (un argument qui confère un nouveau sens à la base *agente*). Ce terme est complet par lui même, puisqu'il est suivi d'une virgule.

Figure 3.3 — Repérage des termes simples faisant partie d'une UTC



Une fois les UTC délimitées, nous les avons ensuite recomposées dans les fichiers Ascii en reliant, pour les besoins du logiciel, par un trait d'union chaque composant de l'UTC. Si nous examinons les contextes qui sont présentés dans la figure 3.3, on s'aperçoit que seul le troisième contexte n'illustre pas une UTC composée du terme *agente*. Dans le trois autres contextes, *agente* se combine au mot qui suit pour former une UTC : *agente biológico*, *agente intercalante* et *agente infectivo*.

Par la suite, ces fichiers ont été régénérés par le logiciel *Makebase* de façon à permettre une nouvelle consultation de la base de données indexée et l'extraction de divers types de listes générées par le programme *Usebase*. À titre d'exemple, nous reproduisons ci-dessous une partie de la liste des UTC du domaine de la biotechnologie des végétaux suivies de leurs contextes.

Figure 3.4 — UTC suivies de leurs contextes

À la fin de cette dernière étape, nous avons enlevé les traits d'union non pertinents qui figuraient dans les UTC et nous avons créé la liste finale des UTC qui contenait environ 2500 UTC non lemmatisées.

3.5 Critères de validation de la terminologie recueillie

Même si l'appartenance de la terminologie au domaine était déjà garantie par le fait qu'elle a été recueillie dans des textes du domaine, nous avons décidé d'être encore plus rigoureuse et d'établir deux autres critères de validation.

a) l'épuisement des types de règles de formation;
b) la vérification dans les dictionnaires.

En effet, ces critères sont d'un tout autre ordre. L'épuisement des types de règles de formation a pour fonction d'établir une limite quant au nombre de données nécessaires à notre recherche. Comme l'un de nos buts était de relever les modes de formation de la terminologie de la biotechnologie de végétaux, nous avons cessé de saisir des données dès que nous n'avons pu trouver de nouveaux modes de formation.

Le critère de vérification dans les dictionnaires sert à valider la terminologie recueillie dans une documentation terminographique et, ainsi, éviter, dans la mesure du possible, d'avoir une terminologie non lexicalisée et instable au niveau sémantique. Étant donné la faible quantité d'ouvrages dictionnairiques en portugais traitant de ce domaine et considérant que la terminologie scientifique a un caractère international en matière de langue, nous avons utilisé plusieurs ouvrages[35] en portugais, en français et en anglais :

a) deux glossaires du domaine en langue portugaise;
b) deux glossaires apparaissant à la fin de livres du domaine en langue portugaise;
c) un dictionnaire multilingue (portugais inclus) du domaine;
d) un lexique des sciences biologiques en français;
e) un dictionnaire du domaine en anglais;
f) trois dictionnaires bilingues (anglais, français);
g) deux banques de terminologie bilingues (anglais, français).

Nous avons également utilisé deux livres du domaine en français qui, bien qu'ils ne soient pas des ouvrages de référence, nous ont aidée pour la validation des données.

Finalement, il faut ajouter à cette liste deux dictionnaires généraux, l'un en langue portugaise et l'autre en français. Ces deux ouvrages ont été utilisés parce qu'on y trouve une fréquence élevée de termes spécialisés.

[35] La référence complète de chaque document figure dans la bibliographie sous la rubrique *Ouvrages de référence*.

Ce dernier critère a été utilisé avec prudence parce que la non-appartenance d'un terme à un dictionnaire n'invalide pas nécessairement son existence, principalement quand il s'agit d'unités complexes formées de plus de deux éléments.

C'est pour cette raison que nous avons conservé certaines UTC qui ne figuraient pas dans les ouvrages terminographiques consultés. Nous avons également retenu les UTC qui n'étaient pas mentionnées intégralement dans les ouvrages de référence, mais pour lesquels nous avons trouvé un ou plusieurs éléments dans les ouvrages en question.

Après avoir appliqué tous les critères mentionnés ci-dessus, nous avons retenu 210 UTC des 2500 UTC initiales. À titre d'illustration, nous présentons, dans le tableau qui suit, les UTC distribuées en fonction de l'identification, dans un ouvrage de référence, des éléments qui les composent.

Tableau 3.1 — Pourcentage d'UTC trouvées dans les ouvrages de référence

Segment repéré dans un ouvrage	**Nbre**	**%**
UTC trouvée au complet dans les ouvrages en portugais	18	8,5
UTC trouvée au complet dans les ouvrages en français ou en anglais	47	22,4
UTC dont une partie figure dans les ouvrages en portugais	33	15,7
UTC dont une partie figure dans les ouvrages en français ou en anglais	77	36,7
UTC non trouvée dans les ouvrages	35	16,7
Total des UTC	210	100

Comme on peut le constater, 83,3 % des UTC figuraient soit au complet soit en partie dans les ouvrages de références consultés. Ces résultats, de même que l'application des autres critères mentionnés auparavant, nous ont permis de nous assurer que les données étaient valables pour notre recherche.

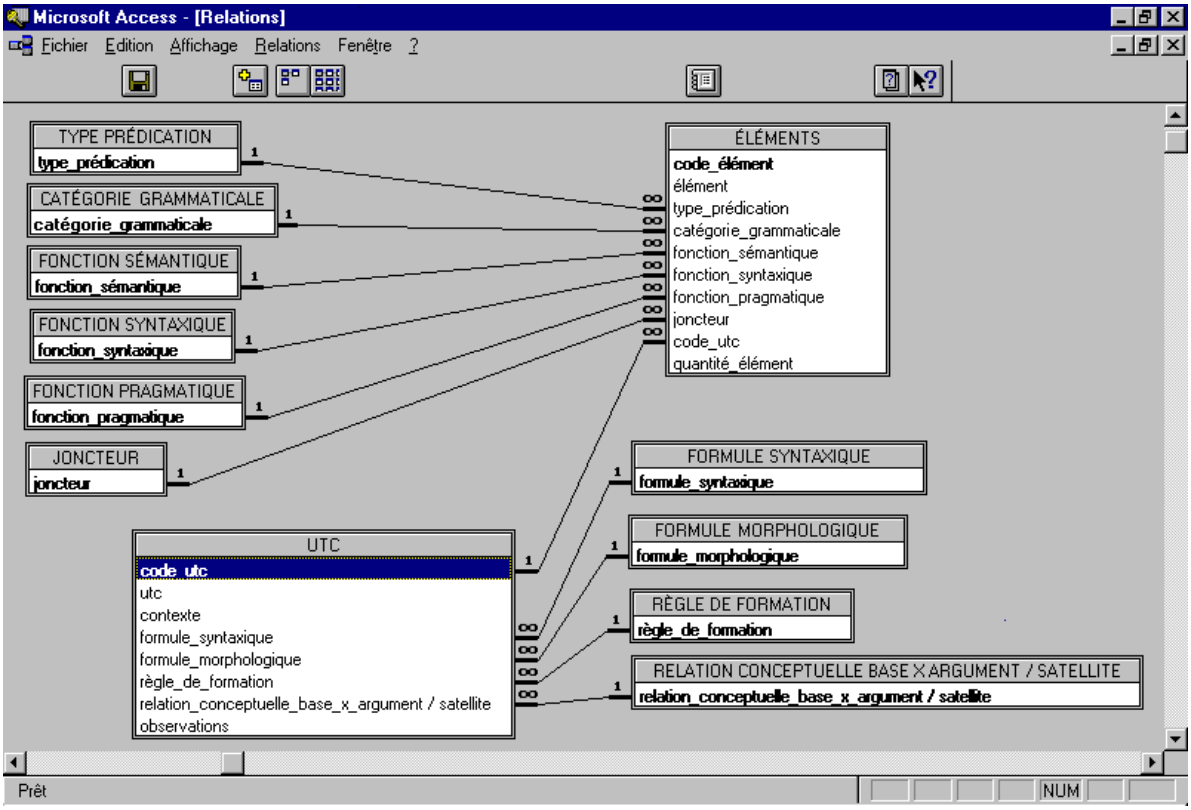
3.6 Systématisation des données

Une fois les données validées, il a fallu les organiser de manière à faciliter la description et l'analyse des UTC. Ainsi, nous avons décidé d'informatiser toute la méthode de la prédication utilisée dans notre recherche. Plus précisément, nous avons automatisé les aspects méthodologiques de manière à effectuer des croisements de données et, par conséquent, à tirer le plus d'avantages possibles des résultats de la recherche. La solution la plus efficace consistait donc à systématiser les données grâce à l'utilisation d'un système de gestion de base de données (SGBD). Parmi les logiciels qui existent, nous avons choisi *Access* (version 2).

3.6.1 Élaboration de la base *Access*

Access est un gestionnaire de base de données relationnel qui tourne sous *Windows* et qui gère les différentes associations entre les données. *Access* comprend plusieurs tables qui se lient les unes par rapport aux autres. Nous présentons ci-dessous un plan des tables et des relations que nous avons définies pour notre corpus de données.

Figure 3.5 — Tables et relations de la base Access



Comme le montre la figure ci-dessus, notre base est composée de onze tables, présentées ici en ordre alphabétique, à savoir : *catégorie grammaticale*, *élément*, *fonction pragmatique*, *fonction sémantique*, *fonction syntaxique*, *formule morphologique*, *formule syntaxique*, *joncteur*, *règle de formation*, *relation conceptuelle base x argument / satellite*, *type de prédication* et *utc*. Chacune de ces rubriques sera définie un peu plus loin dans le chapitre.

Les tables *éléments* et *utc* comportent des champs. Pour la table *éléments*, ces champs sont : *catégorie grammaticale*, *code élément*, *code utc*, *élément*, *fonction pragmatique*, *fonction sémantique*, *fonction syntaxique*, *joncteur*, *quantité d'éléments* et *type de prédication*. Quant à la table *utc*, elle regroupe les champs suivants : *code utc*,

contexte, formule morphologique, formule syntaxique, observations, règle de formation, relation conceptuelle base x argument / satellite et *utc*.

Nous remarquons qu'il y a des tables qui portent des noms de champs identiques. Elles ont alors pour fonction d'établir une association avec la clé de la table à laquelle elles se rapportent. Dans la figure précédente, les clés sont en caractères gras. Nous allons définir les différentes fonctions de l'utilisation des clés dans la section suivante.

Les avantages d'utiliser un gestionnaire de base de données relationnelle pour une recherche de cette nature sont multiples. En ce qui a trait à la mise à jour des données, par exemple, il suffit de corriger un seul élément pour que le système fasse la correction de tous les autres éléments interreliés.

3.6.1.1 Description des tables

Nous avons créé les tables de notre base de données relationnelle en tenant compte de tous les aspects de notre recherche, autrement dit, l'organisation de la base comprend tous les éléments nécessaires pour une analyse systématisée de nos données.

Les tables servent à emmagasiner les données de la base. Chaque table contient un ou plusieurs champs. Dans notre base, il y a des tables qui ont seulement un champ, tandis que d'autres en ont plusieurs. Les tables composées d'un seul champ sont *catégorie grammaticale, fonction sémantique, fonction syntaxique, fonction pragmatique, formule morphologique, formule syntaxique, joncteur, règle de formation, relation conceptuelle base x argument / satellite* et *type de prédication*. Comme nous l'avons vu, les tables *éléments* et *utc* contiennent plus d'un champ. Nous avons réuni dans les deux tableaux ci-dessous la liste des champs de ces deux dernières tables suivies d'une brève explication de leurs contenus et d'exemples d'occurrences possibles s'il y a lieu.

Tableau 3.2 — Table « éléments » et ses champs

Champ	Contenu	Exemples d'occurrence
code_élément	Un code généré automatiquement par le système pour chaque élément d'une UTC.	
élément	Les éléments de chaque UTC.	
type_prédication	L'information sur le type de prédication qui s'applique à l'élément analysé.	étendue - satellite 1 nucléaire - argument1 nucléaire - argument 2 nucléaire - argument 3 nucléaire - base active nucléaire - base statique
catégorie_grammaticale	La catégorie grammaticale de l'élément analysé.	nom, adjectif, adverbe, etc.
fonction_sémantique	La fonction sémantique de l'élément analysé.	agent, appartenance, effet, instrument, localisation, processus, etc.
fonction_syntaxique	La fonction syntaxique de l'élément analysé.	sujet, adjoint de nom, complément du nom, etc.
fonction_pragmatique	La fonction pragmatique de l'élément analysé.	focus, topique
joncteur	Le joncteur d'une UTC.	da, de, do, em, no, etc.
code_utc	Un code généré automatiquement par le système pour chaque UTC.	
quantité_élément	L'information concernant la position de l'élément dans l'UTC.	

Tableau 3.3 — Table « utc » et ses champs

Champ	Contenu	Exemples d'occurrence
code_utc	Un code généré automatiquement par le système pour chaque UTC.	
utc	Les UTC.	
contexte	Les contextes où figurent les UTC analysées.	
règle_de_formation	La règle de formation définie pour chaque UTC.	[Ø [X 1]], [Ø [X 1] [X 2]], [Ø [X 1] [X 2] [X 3]], [Ø [X 1 [X 1/1]]], etc.
formule_morphologique	La formule morphologique de l'UTC analysée.	[N [Adj]], [N [Adj] [Adj]], [N [Adj [Prep + N]]], etc.
formule syntaxique	La formule syntaxique de l'UTC analysée.	[N [SA]], [N [SA] [SA]], [N [SA [SP]]], etc.
relation_conceptuelle_ base_x_argument / satellite	Le rapport conceptuel entre la base et l'argument ou entre la prédication nucléaire et le satellite.	agit sur, contient, se situe dans, résulte de, etc.
observations	Les observations faites au fur et à mesure de l'avancement de la recherche.	

Nous constatons que la plupart des champs de ces deux tables correspondent à une autre table qui porte le même nom, c'est le cas, notamment, du champ *type de prédication*. En vérité, ces champs fonctionnent comme des clés d'accès. Une clé sert à établir des relations entre les tables. Les clés sont, pour nous, le principal avantage d'une base de données relationnelle. Le champ *code utc*, par exemple, est la clé qui fait le lien entre la table *utc* et celle appelée *élément*. La clé permet également la mise à jour rapide et précise des données de plusieurs tables qui ont des rapports entre elles.

Il existe aussi des champs qui n'ont aucune table correspondante et qui ne fonctionnent pas comme des clés. C'est les cas, par exemple, des champs *contexte* et

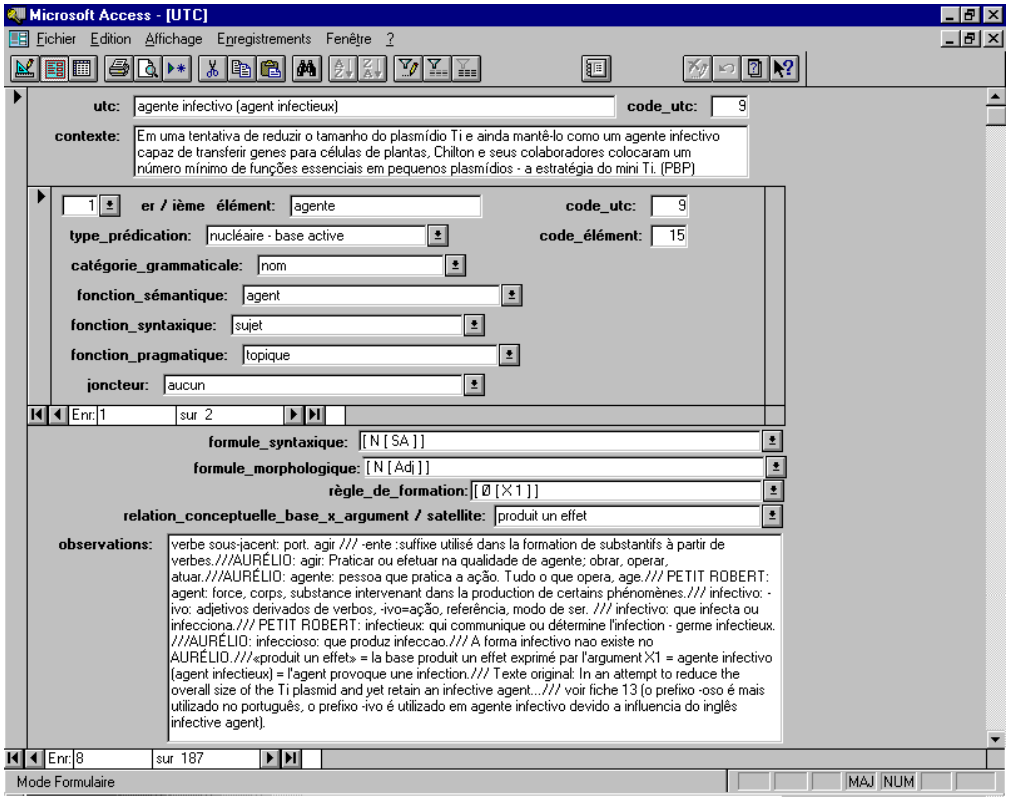
observations. En effet, ces champs ne font pas l'objet d'une table parce qu'ils ne contiennent pas d'informations qui se répètent. En fait, ils sont des champs liés à la table *utc* et portent des informations très variables qui se rapportent uniquement à une seule occurrence de cette table. Toutefois, il est quand même possible de faire des croisements de ces champs avec d'autres champs lorsqu'on prépare une sortie informatique.

Ce type de base permet non seulement de faciliter l'analyse des UTC, mais aussi d'entrer des données de manière systématique. Dans la prochaine section, nous allons traiter des moyens que nous avons utilisés à l'étape de la saisie des données.

3.6.1.2 Formulaire d'entrée des données

À partir des tables que nous avons définies pour notre base, nous avons élaboré un formulaire personnalisé pour l'entrée des données (v. ci-dessous).

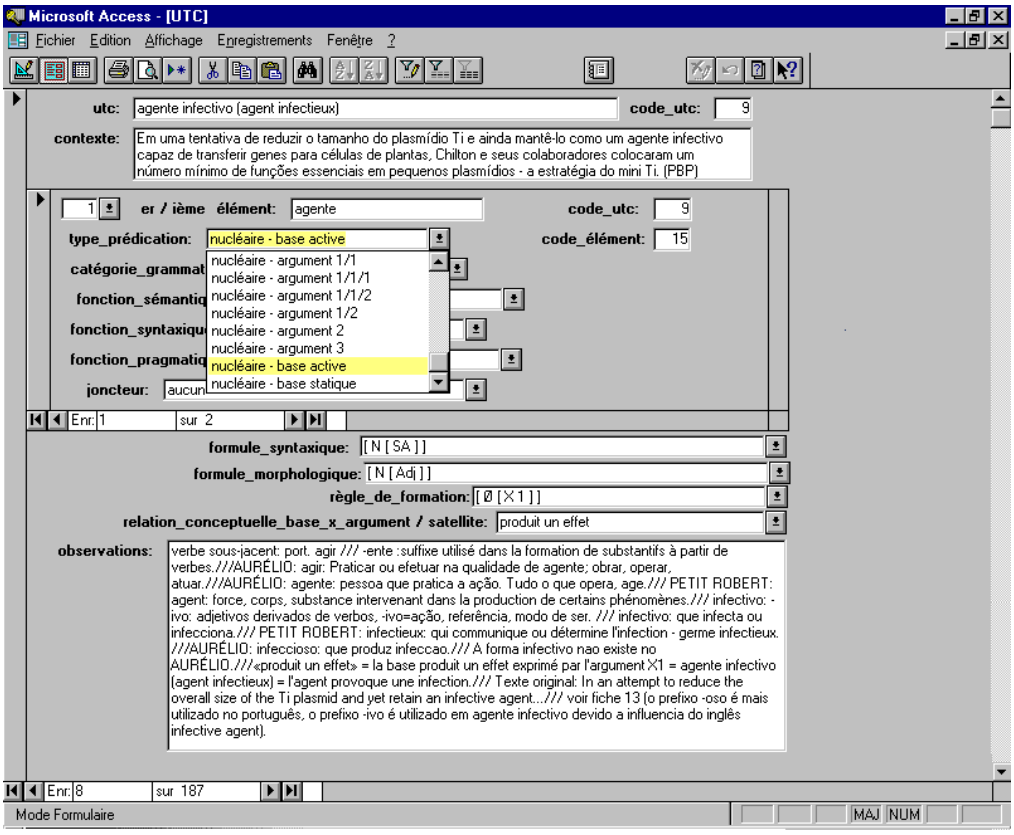
Figure 3.6 — Formulaire d'entrée des données



L'utilisation d'un tel type de formulaire facilite la saisie des données et la rend systématique puisque chaque case du formulaire doit être remplie selon les exigences définies par le système. Cette façon de faire minimise les risques d'erreurs. Toutes les tables qui ont des données fixes comportent des fenêtres où les diverses occurrences sont prédéterminées, ce qui accélère la saisie des données. Dans la figure 3.6, les données fixes comporte une flèche vers le bas à la suite du champ du formulaire. Comme on peut le voir, 11 champs sur 18 fonctionnent de cette façon.

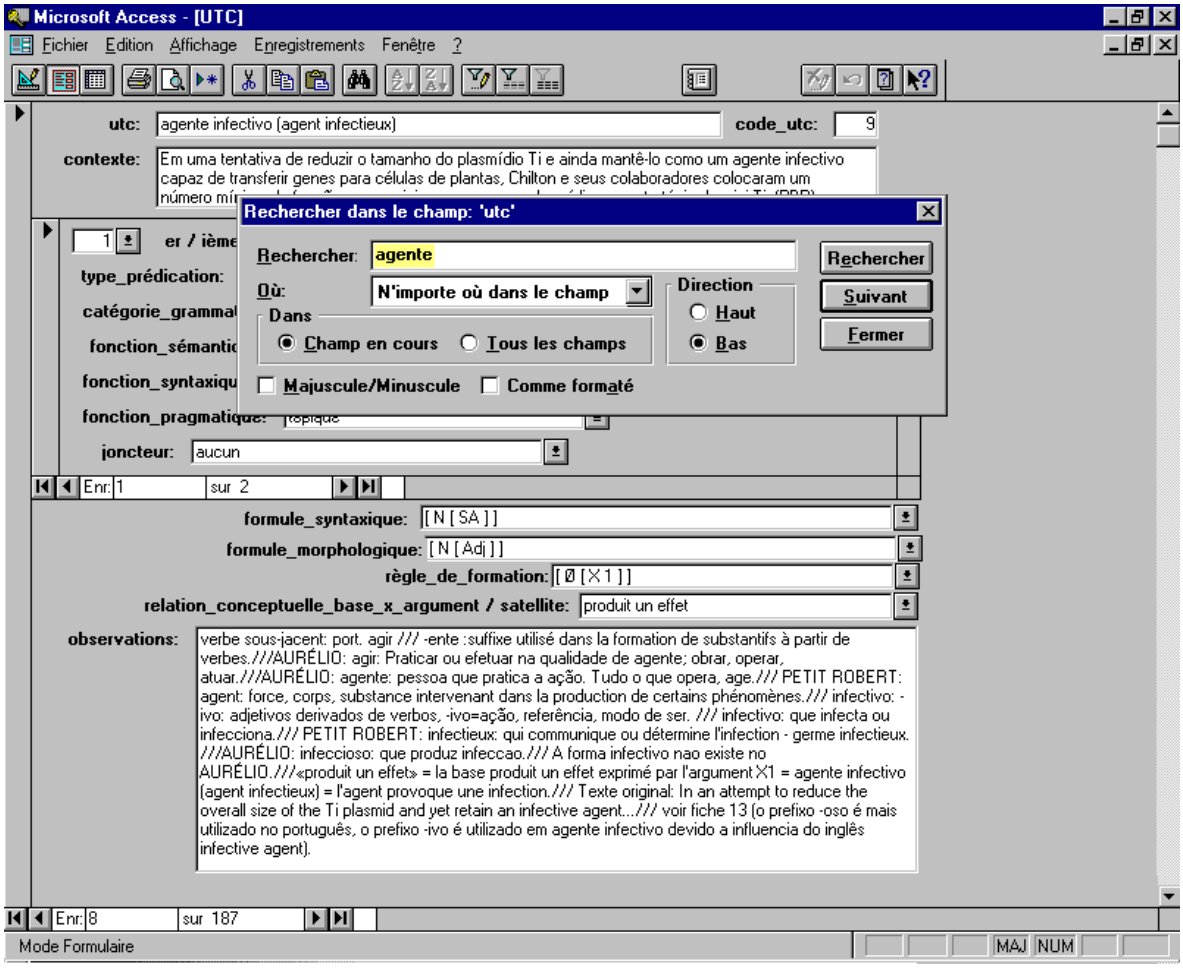
Nous avons reproduit ci-dessous un formulaire où les occurrences de la table *type de prédication* sont listées. On constate ainsi que le champ *type de prédication* peut prendre plusieurs valeurs différentes.

Figure 3.7 — Formulaire d'entrée des données présentant la liste des occurrences de la table « type de prédication »



Pour ce qui est de l'interrogation de la base, elle peut se faire en cherchant, dans n'importe quel champ du formulaire, l'information désirée grâce à divers types d'options. La figure 3.8 présente un exemple de recherche d'éléments. Dans ce cas-ci, le logiciel va repérer toutes les occurrences du terme *agente* qui figurent dans le champ UTC.

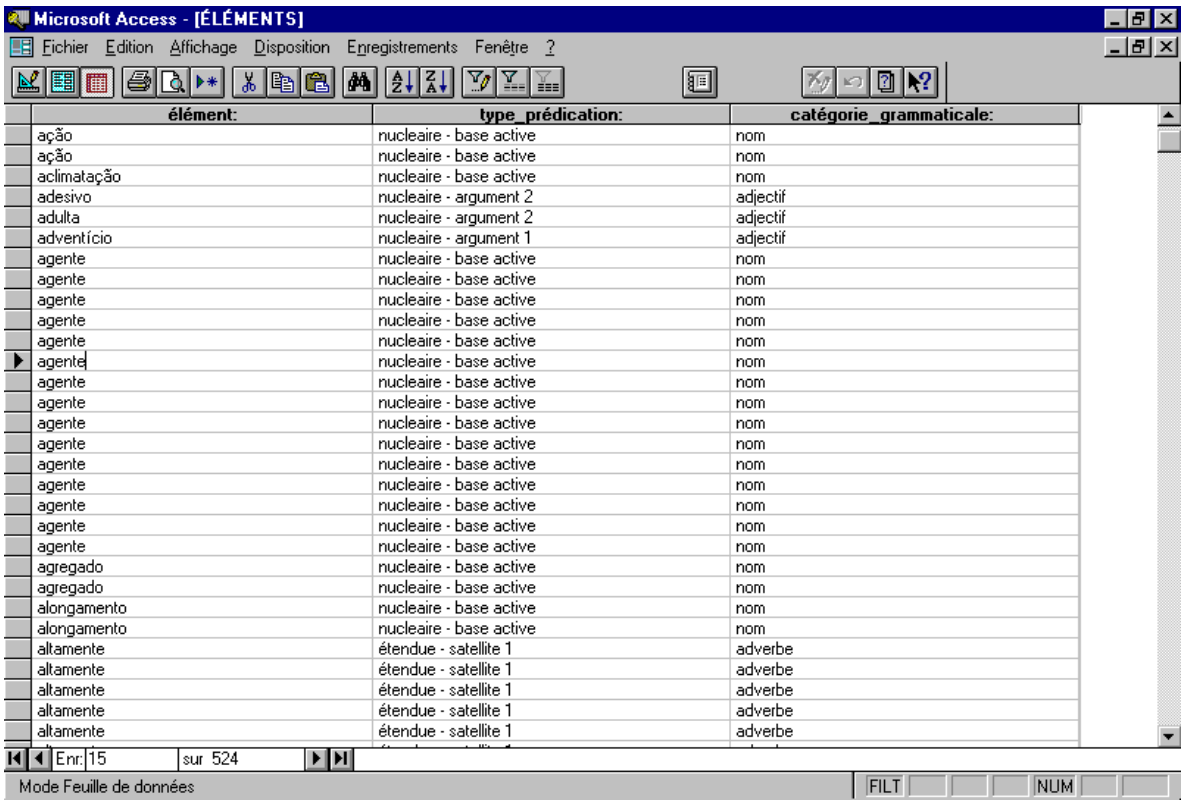
Figure 3.8 — Exemple d'interrogation



Lors de l'entrée des données, il était parfois nécessaire d'obtenir une vue d'ensemble des données déjà insérées. Nous avons alors eu recours à la consultation des données en mode feuille. L'exemple suivant regroupe un ensemble de données des tables *élément*, *type de prédication* et *catégorie grammaticale*. Ces données sont triées par ordre

alphabétique du champ *élément* qui figure dans la première colonne. Il importe de préciser que d'autres tables peuvent être sélectionnées et ordonnées pour répondre aux besoins particuliers de l'utilisateur.

Figure 3.9 — Exemple de données en mode feuille

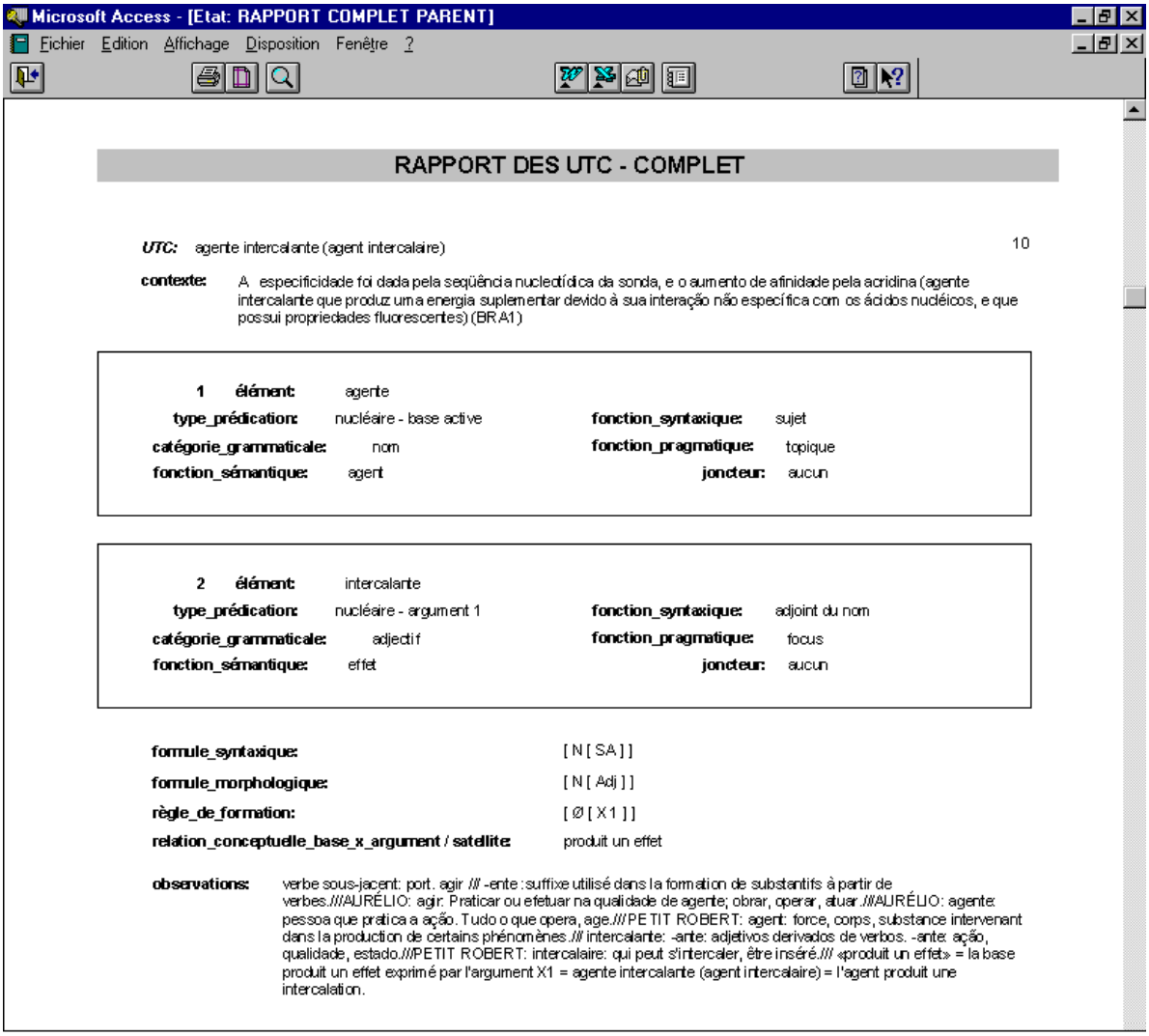


élément:	type_prédication:	catégorie_grammaticale:
ação	nucleaire - base active	nom
ação	nucleaire - base active	nom
aclimatação	nucleaire - base active	nom
adesivo	nucleaire - argument 2	adjectif
adulta	nucleaire - argument 2	adjectif
adventício	nucleaire - argument 1	adjectif
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agente	nucleaire - base active	nom
agregado	nucleaire - base active	nom
agregado	nucleaire - base active	nom
alongamento	nucleaire - base active	nom
alongamento	nucleaire - base active	nom
altamente	étendue - satellite 1	adverbe
altamente	étendue - satellite 1	adverbe
altamente	étendue - satellite 1	adverbe
altamente	étendue - satellite 1	adverbe
altamente	étendue - satellite 1	adverbe

3.6.1.3 Types de sorties générées

En plus d'offrir la possibilité d'imprimer les formulaires et les présentations en mode feuille, *Access* permet également d'imprimer des rapports personnalisés et des graphiques. Nous avons reproduit dans la figure suivante un exemple de rapport qui comprend tous les éléments que nous avons codés pour l'UTC *agente intercalante*. Nous obtenons ainsi, en un seul coup d'oeil, l'ensemble des informations nécessaires pour l'analyse d'*agente intercalante*.

Figure 3.10 — Rapport concernant les UTC



L'avantage d'un rapport de cette nature est de pouvoir choisir quelles données nous voulons présenter et quelle forme visuelle nous voulons leur donner. Il est à noter que l'ensemble des UTC du corpus est présenté de cette façon en annexe.

3.7 Conclusion

Ce chapitre nous a permis d'éclaircir certains points relatifs au repérage de notre corpus et à l'élaboration de la méthode de recherche proprement dite. Nous avons tout d'abord traité du choix de la biotechnologie des végétaux comme domaine pour notre corpus d'analyse. Puis, ce choix a été justifié par des commentaires de nature linguistique et extra-linguistique.

Ensuite, une description de la typologie des textes choisis a été présentée. Cette typologie a été adoptée en tenant compte du public visé par la recherche. Comme nous l'avons d'ailleurs mentionné dans ce chapitre, le domaine de la biotechnologie des végétaux est considéré prioritaire pour le gouvernement brésilien et les pays du *Mercosul*.

Une recherche de cette nature ne pourrait pas être aussi complète sans l'utilisation des outils informatiques. Ainsi, nous avons tiré profit de la performance de quelques logiciels, ce qui nous a permis d'accélérer la collecte tout en effectuant une analyse systématique et cohérente de notre corpus.

En ce qui concerne plus spécifiquement l'analyse des données, l'automatisation a grandement facilité notre tâche étant donné la quantité d'options offertes dans une base relationnelle générée sous *Access*. Ce logiciel s'est avéré très utile pour la présentation des données et leur systématisation puisque le recours à un système qui fait lui-même la gestion de toutes les données ainsi que les diverses associations minimise les risques d'erreurs imputables au traitement de l'information par l'humain.

4. Analyse et interprétation des données

4.1 Introduction

Dans ce chapitre, nous présentons l'analyse et l'interprétation des unités terminologiques complexes du domaine de la biotechnologie des végétaux. À lumière de la grammaire fonctionnelle, nous allons faire ressortir toutes les facettes du comportement de ces structures complexes en utilisant comme point d'appui les règles de formation spécialement conçues à cet effet. Divers aspects de la morphologie, de la syntaxe et de la sémantique, analysés selon une perspective fonctionnaliste, seront utilisés tout au long du chapitre dans le but de faire l'interprétation des UTC de notre corpus.

4.2 Règles de formation

La reconnaissance des modes de formation des UTC employées dans le domaine de la biotechnologie des végétaux a été réalisée au moyen de règles. Ces règles ont été définies selon les principes de base de la structure générale de la prédication de la grammaire fonctionnelle. Le recours aux règles a pour but de systématiser les données tout en rendant l'analyse plus cohérente. En effet, c'est par le biais de ces règles que nous avons pu atteindre le niveau d'abstraction nécessaire pour interpréter les divers processus de formation des UTC.

Avant d'examiner chacune des règles de formation relevées dans notre étude, il faut d'abord rappeler quelques points fondamentaux en ce qui concerne notre conception fonctionnaliste de la structure d'une UTC. Ainsi, nous considérons qu'une UTC est un segment formé d'une *base* suivie d'*arguments* et de *satellites*. La *base* est le noyau de la structure avec lequel les *arguments* ont un rapport direct. Cet ensemble constitué d'une *base* et d'*arguments* forme une *prédication nucléaire*. Les *satellites*, quant à eux, sont des éléments qui ont pour fonction de compléter le concept fourni par les rapports entre la *base* et les *arguments*. De cette façon, les *satellites* ont essentiellement un rapport avec la *base* et ses *arguments* pris comme un tout, c'est-à-dire avec la *prédication nucléaire*. Quant à la jonction de la *prédication nucléaire* et des *satellites*, elle est appelée *prédication étendue*.

Voici comment nous avons conçu notre règle de formation pour illustrer les rapports dont nous venons de traiter.

$$\omega = [\ [\ \phi\ [\ (x_1)\ (x_2)\ \ldots\ (x_n)\ ]\ ]\ [\ (y_1)\ (y_2)\ \ldots\ (y_n)]\ ]$$

Les divers composants d'une UTC, ainsi que l'UTC elle-même, sont représentés par les symboles suivants.

ω = UTC
ϕ = base de l'UTC
x_1, x_2, x_n = arguments
y_1, y_2, y_n = satellites

Les indices 1, 2 et *n* indiquent l'ordre des *arguments* ou des *satellites* dans l'UTC. Les crochets sont utilisés pour cerner les ensembles et les sous-ensembles de rapports établis entre les éléments de l'UTC. Ces rapports peuvent avoir lieu à l'intérieur de la *prédication nucléaire* ou de la *prédication étendue*. Il peut également exister des rapports à l'intérieur d'un ensemble d'*arguments*. Dans ce dernier cas, nous utilisons une barre oblique (/) pour montrer la hiérarchie relationnelle entre les *arguments*. Ce signe

graphique est utilisé essentiellement pour marquer l'expansion d'un *argument*; autrement dit, dans les cas où un *argument* est modifié ou précisé par un autre *argument*. Nous aurons l'occasion de faire ressortir plusieurs exemples de ce type de rapport au cours du présent chapitre.

Nous avons réuni, dans le tableau qui suit, les 10 règles de formation relevées dans notre corpus de données et avons indiqué le pourcentage d'UTC régies par chacune des règles.

Tableau 4.1 — Pourcentage d'UTC par règle de formation

Règle de formation	**Nombre d'UTC par règle de formation**	**%**
[ϕ [x_1]]	100	47,61
[[ϕ [x_1]] [y_1]]	9	4,28
[ϕ [y_1]]	13	6,19
[ϕ [x_1 [$x_{1/1}$]]]	39	18,57
[ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]	10	4,76
[ϕ [x_1] [x_2]]	31	14,76
[ϕ [x_1] [x_2 [$x_{1/2}$]]]	4	1,90
[ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]]]	1	0,47
[ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]]	1	0,47
[ϕ [x_1] [x_2] [x_3]]	2	0,95
Total	210	100

À la lecture de ce tableau, il est possible de constater que certaines règles de formation sont beaucoup plus productives que d'autres. Par ailleurs, si nous considérons le fait qu'un des objectifs de l'analyse de la règle de formation est de nous informer sur le nombre de constituants d'une UTC, nous pouvons connaître, en consultant ce tableau, l'extension des UTC recueillies. Il importe ici de préciser la procédure que nous avons adoptée pour déterminer la quantité d'éléments d'une UTC. En effet, dans cette section introductive, nous avons compté chaque composant sans regarder le rapport de dépendance interne entre les éléments. De plus, les joncteurs n'ont pas fait partie de la somme des composants. Par conséquent, l'UTC *bactéria de vida livre* (bactérie à vie autonome),

désignée par la règle [ϕ [x_1 [$x_{1/1}$]]], a été interprétée comme un segment formé de trois éléments. Nous avons adopté ce type d'analyse afin de dresser un portrait sommaire de l'ensemble des UTC en fonction du nombre d'éléments qui les composent. Il faut toutefois souligner que dans les analyses détaillées des patrons syntaxiques, les joncteurs ainsi que les rapports sémantiques entre les composants ont été pris en compte. Ainsi, en prenant comme point de départ la longueur des UTC et la fréquence d'utilisation des règles de formation, il est possible de faire ressortir quelques tendances du tableau présenté ci-dessus.

D'une part, on constate que beaucoup d'UTC sont formées de deux éléments, soit une *base* suivie d'un *argument* ([ϕ [x_1]]). Ce mode de formation englobe presque la moitié des données (47,61 %). La prépondérance des UTC constituées de deux éléments a déjà été démontrée dans d'autres études portant sur la terminologie de divers domaines. C'est d'ailleurs à cette conclusion que Kocourek ainsi que Daille, pour ne nommer que ces deux auteurs, en sont arrivés. Selon Kocourek, la productivité des formules étendues est très basse. L'auteur attribue ce comportement à la nécessité d'atteindre une communication plus fonctionnelle et mentionne que « [...] La limitation du nombre de mots va ensemble avec la restriction des types de formules syntaxiques des syntagmes lexicaux » (Kocourek, 1991 : 140). Pour sa part, Daille a constaté que les termes formés de deux éléments étaient très fréquents dans le domaine des télécommunications. L'auteure souligne principalement le type *N + ADJ*. qui est sensiblement plus fréquent que les séquences de trois éléments ou plus (v. 1994a).

Si nous ajoutons les UTC formées d'une *base* et de *satellites* représentées par la règle [ϕ [y_1]] aux UTC dont la *base* est suivie d'un *argument* ([ϕ [x_1]]), le pourcentage d'UTC de deux éléments passe de 47,61 % à 53,80 %, ce qui constitue plus de la moitié de l'ensemble des données.

D'autre part, on remarque que les UTC de trois éléments sont caractérisées par trois règles de formation différentes, à savoir :

- [ϕ [x_1] [x_2]];
- [ϕ [x_1 [$x_{1/1}$]]];
- [[ϕ [x_1]] [y_1]].

Si l'on se reporte au tableau 4.1, on s'aperçoit qu'elles représentent 37,61 % de la terminologie recueillie. Ce pourcentage nous semble assez élevé si l'on considère les commentaires de Kocourek et de Daille sur la basse fréquence des séquences très longues (v. *supra*, p. 121).

En ce qui a trait aux UTC qui comportent quatre éléments, elles sont caractérisées par les règles :

- [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]];
- [ϕ [x_1] [x_2 [$x_{1/2}$]]];
- [ϕ [x_1] [x_2] [x_3]].

Toujours en regard du tableau 4.1, seulement 7,61 % des UTC du corpus suivent ces règles. À voir nos données, il semblerait que le domaine de la biotechnologie des végétaux ne soit pas caractérisé par une grande quantité de structures longues. Cette constatation est d'ailleurs renforcée lorsqu'on considère le pourcentage d'UTC composées de cinq éléments (0,94 %). La somme de ces deux types d'UTC (8,55 %) montre bien la faible représentativité des UTC formées de quatre éléments ou plus dans notre domaine d'étude.

D'autres auteurs ont obtenu des résultats similaires dans des domaines qui diffèrent du nôtre. Hoffman mentionne que « Statistical analyses [...] have shown that the frequencies of pattern decrease proportionately to their length, so that the optimum of noun groups, particularly of terminological ones, range from 2 to 4 constituents, which is a compromise between scientific exactitude and language economy » (1987 : 99).

Par ailleurs, Sager et *al* confirment cette tendance quand ils signalent que « Compounds can be made of two, three or more separate elements or compounds as required by the new concept to be represented: in practice five- or six-element compounds are rare, but variations exist between special subjects » (1980 : 272).

Dans les pages qui suivent, nous allons approfondir notre analyse en nous penchant sur les rapports morphologiques, syntaxiques et sémantiques de ces UTC. En effet, pour chaque règle de formation, l'ensemble des UTC comporte diverses caractéristiques particulières qui distinguent une UTC d'une autre, constituant ainsi des sous-ensembles de données. Ces subdivisions résultent de la perception, selon divers angles d'analyse, du corpus complet. Ces angles constituent, en fait, les paramètres que nous avons imposés à la description, à l'analyse et à l'interprétation de nos données.

Afin de faciliter la compréhension, nous avons décidé de systématiser la présentation des paramètres d'analyse établis pour chaque règle de formation relevée dans notre corpus. Ainsi, nous allons interpréter les UTC selon

- leurs formules syntaxiques;
- leurs formules morphologiques;
- leurs relations conceptuelles *base* x *argument* / *satellite*.

Pour y arriver, il faut également examiner les diverses parties des UTC. Par conséquent, nous avons créé six paramètres d'analyse pour chaque composant de l'UTC :

- le type d'élément;
- la catégorie grammaticale;
- la fonction sémantique;
- la fonction syntaxique;
- la fonction pragmatique;

- le joncteur.

Il importe de préciser que cette méthode sera adoptée pour l'interprétation des données de chaque règle de formation, mais que pour éviter les redondances seuls les paramètres qui distinguent une règle de l'autre seront expliqués en détail.

4.2.1 [ϕ [x_1]]

La règle de formation [ϕ [x_1]] représente les UTC du type *base* + *argument*. Elle est formée essentiellement d'une prédication nucléaire puisque l'élément de droite est un argument et non un satellite. Comme nous l'avons déjà précisé, il s'agit d'une règle très productive dans le domaine de la biotechnologie des végétaux.

L'analyse de cette règle a été divisée en deux sections. La première traite de la *base* (ϕ) de l'UTC et la deuxième de l'*argument* (x_1). Pour ces deux composants, nous ferons les interprétations qui s'imposent suivant les paramètres d'analyse définis dans la section antérieure.

4.2.1.1 Base

Lorsqu'on analyse les constituants de ce type d'UTC, on note que la base est essentiellement *statique* (61 %) comme le montre le tableau 4.2.

Tableau 4.2 — Pourcentage d'UTC par type de base pour la règle [ϕ [x_1]]

Type de base	**Nombre d'UTC par type de base**	**%**
active	39	39,00
statique	61	61,00
Total	100	100

Une base est dite *active* lorsqu'elle dérive d'un verbe, c'est-à-dire quand un verbe est sous-jacent à sa forme nominale. Dans les autres cas, la base est dite *statique*.

Les *bases active*s sont produites par dérivation déverbale. Il s'agit donc de nominalisations de la forme verbale. Guilbert avait déjà fait référence à ce type de base quand il a analysé ce qu'il nommait le *composé à base verbale*. Selon lui,

> Dans le composé à base verbale [...] le passage de la phrase sous-jacente au composé réalisé, consiste dans la transformation d'une séquence phrastique (SN + SV [V + SN]) en un syntagme nominal. La réduction de la forme de phrase à celle de nom [...] a pour résultat d'enlever aux constituants de la phrase leur liberté de variation, et, surtout, de transformer le caractère du prédicat verbal qui, d'actuel dans la phrase, devient virtuel dans le composé nominal. Cette transformation se manifeste notamment par la forme de la base verbale du composé dépourvue de toute marque syntaxique de morphème libre (personne, nombre), réduite à l'état de signe verbal neutralisé, comparable à la forme infinitive du verbe, puisqu'elle exprime la valeur verbale abstraite pourvue de la seule capacité de rection d'un complément (1975 : 262).

Toutefois, déterminer l'origine d'une base, n'est pas une mince tâche. Parfois, il est difficile de définir quelle forme a été créée en premier. Est-ce le nom qui est dérivé du verbe ou le contraire? Normalement les noms qui expriment une action ou un processus sont dérivés de verbes. Prenons par exemple les bases *ação* (action), *alongamento* (élongation), *amplificação* (amplification), *aplicação* (application), *associação* (association), *desenvolvimento* (développement), *indução* (induction), *iniciação* (initiation),

propagação (propagation) et *reinfecção* (réinfection) qui ont été classées comme des *bases actives* dans les UTC suivantes :

- ***ação** endógena* (action endogène);
- ***alongamento** celular* (élongation cellulaire);
- ***amplificação** gênica* (amplification génique);
- ***aplicação** biotecnológica* (application biotechnologique);
- ***associação** simbiótica* (association symbiotique);
- ***desenvolvimento** foliar* (développement de la feuille);
- ***indução** floral* (induction florale);
- ***iniciação** floral* (initiation florale);
- ***propagação** clonal* (propagation clonale);
- ***reinfecção** viral* (réinfection virale).

Il s'agit de bases qui expriment toutes un processus. Ainsi une ***indução** floral* (induction florale) est un processus de « stimulation des phénomènes physiologiques qui président à la floraison » (Quirion et Bourbeau, s.d. : 359). Toutefois, une base qui a pour fonction sémantique un processus peut également être une *base statique* comme le montrent les deux exemples suivants :

- ***cultura** de anteras* (culture d'anthères);
- ***embriogênese** somática* (embryogenèse somatique).

L'emploi de *bases active*s est un phénomène qui découle du comportement du discours scientifique et technique, tandis que l'utilisation de structures verbales marque les textes de type narratif. En effet, la formation du nom à partir de verbes constitue un recours qui vise à répondre aux exigences syntaxiques et sémantiques des textes scientifiques et techniques. Il s'agit de l'expression du verbe par le biais de la nominalisation. L'utilisation de la nominalisation vient pallier la déficience de la structure verbale qui ne permet pas les

concaténations multiples dans le texte. De plus, la structure nominale aide à mieux exprimer les idées, les faits et les phénomènes qui caractérisent le discours terminologique.

La reconnaissance des *bases actives* facilite l'organisation des contraintes lors de l'établissement des fonctions sémantiques. En effet, l'analyse des verbes sous-jacents à la forme nominale fournit des indices sur le rôle sémantique joué par les arguments de l'UTC. Autrement dit, le fait de connaître l'action exprimée par le verbe sous-jacent à la base permet de mieux définir le processus ou l'entité associée au processus. Prenons par exemple les UTC de *base active* telles que :

- ***alongamento*** *celular* (élongation cellulaire);
- ***reinfecção*** *viral* (réinfection virale).

Dans le premier cas, le verbe sous-jacent de *alongamento* (élongation) est *alongar* (allonger). Dans la phrase, *Pedro alongou o braço para pegar o copo* (Pierre a allongé[36] le bras pour prendre le verre), nous pouvons dire que ce verbe demande un agent (*Pedro*) et un patient (*braço*). Dans le cas de l'UTC ***alongamento*** *celular* (élongation cellulaire), l'argument *celular* (cellulaire) remplit la même fonction sémantique que *braço* (bras) et non que *Pedro* (Pierre). Ainsi, *celular* (cellulaire) est le patient du processus et non l'agent.

Dans la seconde UTC, *viral* (virale) remplit la fonction d'*agent* puisque dans cette UTC c'est le virus qui réinfecte. Dans la phrase *A bactéria reinfectou a garganta* (La bactérie a réinfecté la gorge), on constate que le verbe sous-jacent *reinfectar* (réinfecter) demande un agent (*bactéria*) et un patient (*garganta*). En comparant ces fonctions sémantiques avec le rôle de l'argument *viral* (virale) dans l'UTC citée, on remarque que ce dernier est un *agent*. En effet, c'est la nature sémantique proprement dite du mot virus,

[36] Nous avons utilisé le verbe allonger au lieu d'élonger puisque l'emploi de ce dernier est considéré comme vieilli.

laquelle fait appel à un micro-organisme qui infecte, qui permet d'affirmer que l'argument a la fonction agent.

Toutefois, il faut préciser que cette démarche n'est pas valide dans tous les cas. Dans les UTC ***aplicação*** *biotecnológica* (application biotechnologique) et ***associação*** *simbiótica* (association symbiotique), les arguments ne possèdent ni la fonction d'*agent* ni celle de *patient*. Nous reviendrons sur ce sujet un peu plus loin quand nous traiterons de la fonction sémantique des arguments.

En principe, la reconnaissance des suffixes aide d'une part à découvrir le changement d'une catégorie et, d'autre part, à établir la fonction sémantique du composant de l'unité terminologique.

En ce qui concerne la formation suffixale des *bases actives*, plus de la moitié (51,28 %) d'entre elles comportaient le suffixe *-ção* comme l'indique le tableau qui suit.

Tableau 4.3 — Pourcentage d'UTC de base active par suffixe pour la règle [ϕ [x_1]]

Suffixe	**Nombre d'UTC de base active par suffixe**	**%**
-ção	20	51,28
-ente	11	28,20
-ado	2	5,13
-mento	2	5,13
-agem	1	2,56
-dor	1	2,56
-ância	1	2,56
ø[37]	1	2,56
Total	39	100

[37] La catégorie ø a été employée pour l'UTC *ganho genético* (gain génétique). La base *ganho* (gain) est un participe irrégulier du verbe ganhar (gagner), dont la variante *ganhado* (gagné) est le participe formé du suffixe *-ado.*

Il faut remarquer que les suffixes *-ção*, *-mento*, *-agem*, *-ância* sont utilisés dans le corpus pour former des *bases active*s dont la fonction sémantique est un *processus*. Pour ce qui est des suffixes *-ente*, et *-dor*, ils remplissent la fonction sémantique *agent*, c'est-à-dire une entité associée à un processus. Nous avons réuni quelques exemples dans le prochain tableau.

Tableau 4.4 — Suffixes et fonctions sémantiques des bases actives pour la règle [ϕ [x_1]]

Suffixe	**Fonction sémantique**	**Exemple d'UTC**
-ção	*processus*	*hibrida**ção** somática* (hybridation somatique) *propaga**ção** clonal* (propagation clonale)
-mento	*processus*	*alonga**mento** celular* (élongation cellulaire) *desenvolvi**mento** foliar* (développement de la feuille)
-agem	*processus*	*clon**agem** rápida* (clonage rapide)
-ância	*processus*	*domin**ância** apical* (dominance apicale)
-ente	*agent*	*ag**ente** mutagênico* (agent mutagène) *ag**ente** patogênico* (agent pathogène)
-dor	*agent*	*marca**dor** genético* (marqueur génétique)

Plus spécifiquement, si on analyse l'UTC *clon**agem** rápida* (clonage rapide), qui utilise le suffixe *-agem*, on note que la *base* est *active* puisque le verbe sous-jacent est *clonar* (cloner). Si on consulte le *Novo dicionário da língua portuguesa*, on constate que le nom *clone* (clone) est apparu dans la langue avant le nom *clonagem* (clonage). En fait, ce dernier est créé à partir du nom *clone* (clone) et du suffixe *-agem*. Ce qui est logique, puisque avant d'avoir le processus de clonage, il faut absolument avoir au moins l'idée de clone. Un processus (p. ex. clonage) est effectué avec, par ou pour une entité (p. ex. clone). En d'autre mots, il y a tout d'abord la chose ou l'idée de *clone* (clone) pour ensuite avoir l'action *clonar* (cloner) produite par le processus *clonagem* (clonage). Cette même analyse s'applique à *hibridação*[38] *somática* (hybridation somatique).

[38] *híbrido* (hybride) › *hibridar* (hybrider) › *hibridação* (hybridation).

La prédominance des *bases statique*s (61 %) réaffirme le caractère nominatif du discours terminologique. Par rapport à la provenance du vocabulaire utilisé par ce type de base, il faut remarquer qu'il semble y avoir une distinction entre la source lexicale des *bases actives* et celle des *bases statiques*.

Nous avons observé que, d'une part, les *bases active*s correspondent principalement à des signes linguistiques provenant d'un type de vocabulaire fonctionnel (terme emprunté à Guilbert) présent dans la terminologie propre à plusieurs domaines. Ces bases ont souvent besoin d'arguments pour préciser leur signification. Cette tendance qui se dégage de notre corpus rejoint ce que Guilbert avait déjà constaté dans son étude portant sur le domaine de l'aviation. En effet, il a remarqué que ce type de bases (caractérisées par l'utilisation des mots fourre-tout tels que machine, système et agent) constitue « des sortes de lexèmes fonctionnels de tout vocabulaire technique appelant par leur signification imprécise une complémentation et une adaptation à chaque champ sémantique particulier » (1965 : 269).

D'autre part, nous avons constaté que les *bases statiques* représentent surtout des signes linguistiques spécifiques du domaine de la biotechnologie des végétaux ou d'une langue de spécialité connexe. D'ailleurs, Guilbert a observé le même phénomène dans son corpus quand il a analysé le rôle des bases dans la formation des syntagmes nominaux à vocation lexicale (v. *ibid*.) Une autre différence entre ces deux types de base est que les *bases actives* font référence à un concept moins concret tandis que les *bases statiques* renvoient à une notion sémantique plus concrète.

Les deux prochains tableaux illustrent le phénomène que nous venons d'analyser. Le tableau 4.5 présente des exemples de *bases statiques* et de *bases actives* qui utilisent une terminologie fonctionnelle ou une terminologie spécifique au domaine. Dans le tableau 4.6, nous reprenons les mêmes exemples pour mettre en lumière le caractère plus ou moins concret des bases utilisées dans ces UTC.

Tableau 4.5 — Terminologie employée selon le type de base pour la règle [ϕ [x_1]]

UTC	Base statique	Base active	Terminologie fonctionnelle	Terminologie du domaine ou d'un domaine connexe
***ação** enzimática* (action enzymatique)		x	x	
***alteração** cromossômica* (altération chromosomique)		x	x	
***desenvolvimento** foliar* (développement de la feuille)		x	x	
***genoma** nuclear* (génome nucléaire)	x			x
***híbrido** somático* (hybride somatique)	x			x
***raiz** adventícia* (racine adventice)	x			x
***tecido** embrionário* *(tissu embryonnaire)*	x			

Tableau 4.6 — Notion sémantique employée selon le type de base pour la règle [ϕ [x_1]]

UTC	Base statique	Base active	Notion plus concrète	Notion moins concrète
***ação** enzimática* (action enzymatique)		x		x
***alteração** cromossômica* (altération chromosomique)		x		x
***desenvolvimento** foliar* (développement de la feuille)		x		x
***genoma** nuclear* (génome nucléaire)	x		x	
***híbrido** somático* (hybride somatique)	x		x	
***raiz** adventícia* (racine adventice)	x		x	
***tecido** embrionário* *(tissu embryonnaire)*	x		x	

Pour ce qui est de la catégorie grammaticale, toutes les bases sont des *noms* puisque notre étude se restreint aux UTC nominales. Quant aux fonctions syntaxiques, elles remplissent toujours la fonction de *sujet*. Il en va de même de la fonction pragmatique qui assume invariablement le rôle de *topique* pour ce qui est des bases des UTC. Il ne pourrait pas en être autrement étant donné que la fonction pragmatique *topique* est attribuée au composant de l'UTC auquel tous les arguments et les satellites font référence dans le but de préciser, de décrire ou d'expliquer l'objet de la communication.

Quant à la fonction syntaxique, il faut souligner que, suivant les principes de la grammaire fonctionnelle, le *sujet* est défini comme l'élément clé pour déterminer le comportement syntaxique des composants d'une UTC. Ainsi, d'après la grammaire fonctionnelle, l'analyse syntaxique d'une UTC est faite selon la perspective du composant qui remplit la fonction *sujet*. Prenons par exemple l'UTC *célula somática* (cellule somatique), la base *célula* (cellule) est le *sujet* puisque c'est sur elle qu'on veut donner de l'information quant à la caractéristique désignée par l'argument *somática*[39] (somatique). C'est la *célula* (cellule) qui est *somática* (somatique) et non le contraire. La définition fonctionnelle de *célula somática* (cellule somatique) est « cellule du corps d'un organisme, par opposition aux cellules germinales, ovules et spermatozoides » (v. Termium). Par conséquent, la cellule appartient au corps, au soma.

Les sujets des UTC peuvent combler plusieurs fonctions sémantiques. D'ailleurs, ce phénomène a déjà été observé par Dik qui mentionne que « It will be clear that the perspectivizing functions Subj/Obj cannot be reduced to semantic functions. For the whole point of Subj and Obj is that they may be assigned to terms with different semantic functions, thus reorganizing the basic orientation inherent in the predicate frame » (1989 : 216).

[39] Relatif au soma, au corps.

Pour ce qui est des UTC régies par la règle [ϕ [x_1]], leurs composants, qui remplissent la fonction syntaxique de *sujet*, peuvent acquérir différentes fonctions sémantiques, telles que *agent*, *effet*, *entité*, *groupement*, *localisation* ou *processus*. La répartition des fonctions sémantiques des *bases-sujets* des UTC est d'ailleurs présentée dans le tableau 4.7.

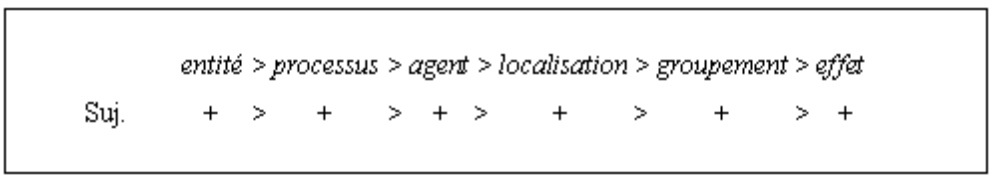
Tableau 4.7 — Pourcentage des fonctions sémantiques des bases-sujets pour la règle [ϕ [x_1]]

Fonction sémantique de la base-sujet	**Nombre de base-sujet par fonction sémantique**	**%**
agent	17	17,00
*effe*t	1	1,00
entité	41	41,00
groupement	3	3,00
localisation	7	7,00
processus	31	31,00
Total	100	100

Il ressort de ce tableau que les fonctions sémantiques *entité* (41 %) et *processus* (31 %) prédominent. Ensemble, elles englobent 72 % des fonctions.

En suivant le principe de la grammaire fonctionnelle, selon lequel la syntaxe est analysée au moyen de la sémantique, nous pouvons reprendre le schéma produit par Dik sur la hiérarchie de la fonction sémantique (v. *supra* p. 21) et l'adapter à nos données. Cette hiérarchie est représentée dans la figure 4. 1.

Figure 4.1 — Hiérarchie de la fonction sémantique des bases-sujets pour la règle [ϕ [x_1]]



La hiérarchie présentée dans la figure 4.1 fait ressortir le niveau d'importance de la fonction sémantique illustrée pour la fonction syntaxique *sujet*. De cette façon, en ce qui concerne les *bases-sujets*, la fonction *entité* est plus importante, du point de vue de la fréquence d'utilisation, que la fonction *processus* et ainsi de suite.

Afin d'illustrer l'emploi de ces fonctions sémantiques, nous avons regroupé des exemples de notre corpus dans le tableau ci-dessous.

Tableau 4.8 — Fonction sémantique des bases pour la règle [ϕ [x_1]]

Fonction sémantique de la base	**UTC**
entité	***antera*** *madura* (anthère mature) ***casa*** *de vegetação* (serre) ***híbrido*** *somático* (hybride somatique) ***raiz*** *adventícia* (racine adventice)
processus	***hibridação*** *somática* (hybridation somatique) ***mutação*** *somática* (mutation somatique) ***propagação*** *clonal* (propagation clonale)
agent	***agente*** *patogênico* (agent pathogène) ***marcador*** *genético* (marqueur génétique) ***vetor*** *genético* (vecteur génétique)
localisation	***axila*** *foliar* (aiselle de la feuille) ***sítio*** *de restrição* (site de restriction)
groupement	***agregado*** *celular* (agrégat cellulaire) ***variedade*** *semi-anã* (variété semi-naine)
effet	***efeito*** *inibitório* (effet inhibiteur)

4.2.1.2 Argument

Passons maintenant à l'analyse de l'argument x_1 de la règle [ϕ [x_1]]. Cet argument occupe la deuxième place dans la structure et se situe à côté de la base de l'UTC. Il est considéré comme un argument parce qu'il a des rapports avec la base de l'UTC et non avec la prédication nucléaire comme c'est le cas des satellites.

Ces arguments sont pour la plupart des *adjectifs*, mais quelques *syntagmes prépositionnels* ont également été relevés. Nous présentons, dans le tableau qui suit, la répartition des catégories grammaticales de ces arguments.

Tableau 4.9 — Pourcentage de catégories grammaticales par argument pour la règle [ϕ [x_1]]

Catégorie grammaticale	**Nombre d'arguments par catégorie grammaticale**	**%**
adjectif	81	81,00
syntagme prépositionnel	19	19,00
Total	100	100

Les arguments de type syntagmes prépositionnels sont composés selon la séquence *Prép. + N*. Le type de prépositions peut varier, c'est toutefois le joncteur *de* qui prédomine. Voici quelques exemples tirés de notre corpus :

- *agregado* ***de*** *células* (agrégat de cellules);
- *ambiente* ***de*** *cultura* (milieu de culture);
- *casa* ***de*** *vegetação* (serre);
- *célula* ***em*** *suspensão* (cellule en suspension);
- *cultura* ***de*** *anteras* (culture d'anthères);
- *cultura* ***em*** *batelada* (culture en batch);
- *híbrido* ***por*** *complementação* (hybride par complémentation);
- *mapa* ***de*** *restrição* (carte de restriction)
- *mecanismo* ***de*** *regulação* (mécanisme de régulation);
- *método* ***de*** *propagação* (méthode de propagation);
- *propagação* ***em*** *massa* (propagation en masse);
- *sítio* ***de*** *restrição* (site de restriction);
- *vetor* ***de*** *clonagem* (vecteur de clonage).

Ces exemples font ressortir l'importance du choix de la préposition. Comme nous l'avons vu au chapitre 1 (v. supra, p. 19), la préposition a une valeur telle qu'elle confère à l'UTC un sens précis qu'elle ne pourrait avoir autrement. En effet, la préposition a pour rôle d'établir une hiérarchie fonctionnelle entre le premier élément (la base) et le deuxième (l'argument). Dans notre étude, nous analysons la préposition en même temps que l'argument, sans toutefois oublier qu'elle est, d'une certaine manière, exigée par la base de l'UTC. À ce sujet, Cervoni mentionne, dans une analyse syntactico-sémantique de la préposition en français, que « [...] en tout emploi la préposition est plus à son régime (ou terme B) qu'à l'élément auquel elle le rattache (terme A); morphème "séparable", elle ne l'est pas au point de ne pas avoir d'attache privilégiée avec l'élément nominal dont elle complète la morphologie intégrée » (1991 : 94).

Sur le plan syntactico-sémantique, la préposition est un élément essentiel duquel dépend considérablement le sens de l'UTC. En effet, la préposition matérialise un rapport sémantique entre deux composants d'une UTC. Le changement de la préposition peut parfois occasionner un changement de la notion exprimée par l'UTC.

En analysant de près la préposition *de*, nous remarquons, au-delà de sa fréquence élevée, qu'elle agit comme un opérateur d'interprétations sémantiques entre les éléments de l'UTC. Cervoni constate que « Comme la préposition *de*, considérée comme la plus abstraite des prépositions françaises, est celle qui, du fait même de son abstraction, est la plus polyvalente, ses latitudes d'emploi paraissent presque illimitées » (*id*. : 14). Une analyse similaire pourrait être effectuée pour le *de* en langue portugaise.

Examinons les UTC *agregado* ***de*** *células* (agrégat de cellules), *ambiente* ***de*** *cultura* (milieu de culture), *cultura* ***de*** *anteras* (culture d'anthères), *mapa* ***de*** *restrição* (carte de restriction), *sítio* ***de*** *restrição* (site de restriction) et *vetor* ***de*** *clonagem* (vecteur de clonage).

La préposition *de* exprime différentes relations. Elle marque la *contenance* dans *agregado* ***de*** *células* (agrégat de cellules), *ambiente* ***de*** *cultura* (milieu de culture) et *mapa* ***de*** *restrição*[40] (carte de restriction). Dans le cas de *cultura* ***de*** *anteras* (culture d'anthères), la préposition indique la *production d'un résultat* tandis que dans *vetor* ***de*** *clonagem* (vecteur de clonage) elle représente la *production d'un effet*.

La fonction sémantique de chaque composant de l'UTC qui comporte une préposition a sûrement une incidence sur cette diversité d'interprétations. Dans *agregado* ***de*** *células* (agrégat de cellules), nous avons un rapport entre une fonction sémantique de *groupement* exprimée par la base *agregado* (agrégat) et une de *contenance* représentée par *células* (cellules). Dans *vetor* ***de*** *clonagem* (vecteur de clonage), la base *vetor* (vecteur) est un *agent* tandis que l'argument *clonagem* (clonage) est un *effet*.

C'est donc l'interprétation du rôle ou de la fonction sémantique de chaque élément à droite et à gauche du joncteur qui va définir la valeur sémantico-fonctionnelle de la préposition. Le fait de rattacher la préposition avec l'argument provoque des adaptations et des restrictions sémantiques au moment de la combinaison. La reconnaissance de ces transformations aide à mettre en valeur les rapports syntactico-fonctionnels établis dans la formation des UTC. Toutefois, le composant à droite de la préposition est lui aussi important en tant qu'élément avec lequel l'argument réalise le rapport pour constituer le concept généré par la prédication.

Ces observations s'avèrent essentielles quand on pense à un processus d'automatisation. Le repérage automatique des prépositions ne pose pas vraiment de problèmes puisqu'elles constituent un ensemble fermé dans une langue. Toutefois, quand on veut que le repérage d'information dépasse le simple dépistage dans un contexte, la

[40] Dans le cas de cette UTC le rapport est de *contenance* puisque c'est une carte que contient des sites de restriction. En réalité, il y a eu un effacement de l'élément « sítios » de l'argument *de restrição*. À l'origine, cette UTC devait être *mapa* ***de*** *sítios de restrição* (carte de sites de restriction).

problématique se présente tout autrement. À ce sujet Ritzke signale quelques difficultés provenant d'une analyse automatique des prépositions. Selon lui,

> [...] comme les éléments prépositionnels se montrent particulièrement chargés de fonctions différentes en ce qui concerne leur syntaxe et leur sémantique, c'est ici que les difficultés d'une analyse automatique commencent. Très souvent des fonctions différentes sont exprimées par une forme identique, de sorte que l'identification des valeurs précises n'est plus possible en exploitant seulement la forme graphématique et l'information catégorielle. Car la même préposition peut introduire des syntagmes prépositionnels aux fonctions différentes [...] (1981 : 142).

Pour ce qui est de la fonction syntaxique des arguments, nous en avons observé trois types, à savoir l'*adjoint de nom*, le *complément du nom* et le *complément circonstanciel*. La répartition de ces fonctions selon les arguments est illustrée dans le tableau qui suit.

Tableau 4.10 — Pourcentage d'arguments par fonction syntaxique pour la règle [ϕ [x_1]]

Type de fonction syntaxique	**Nombre d'arguments par fonction syntaxique**	**%**
adjoint de nom	91	91,00
complément du nom	4	4,00
complément circonstanciel	5	5,00
Total	100	100

L'établissement de ces fonctions est fondé sur l'organisation et sur le rôle des composants de la structure interne de l'UTC. Ainsi, un argument a la fonction syntaxique d'*adjoint de nom* quand il a une valeur adjectivale, de *complément du nom* quand il est introduit par une préposition et qu'il n'a pas une valeur adverbiale et de *complément circonstanciel* s'il est prépositionnel et a une valeur adverbiale.

Lorsqu'on consulte le tableau 4.10, on remarque que la grande majorité des arguments de la règle [ϕ [x_1]] sont des *adjoints de nom*. Ce résultat découle

principalement de la fréquence élevée d'arguments appartenant à la catégorie grammaticale des adjectifs (v. tableau 4.9).

Les cinq cas de fonction de *complément circonstanciel* figurent ci-dessous.

- *célula* ***em suspensão*** (cellule en suspension);
- *cultura* ***em batelada*** (culture en batch);
- *cultura* ***em suspensão*** (culture en suspension);
- *híbrido* ***por complementação*** (hybride par complémentation);
- *propagação* ***em massa*** (propagation en masse).

Ces *compléments circonstanciels* sont attribués à des syntagmes prépositionnels marqués essentiellement par la fonction sémantique *mode*, sauf pour ce qui est de l'UTC *propagação* ***em massa*** (propagation en masse). Dans ce dernier cas, la fonction sémantique est plutôt d'*intensité* même si, à première vue, on pourrait penser qu'il s'agit d'un *mode*. En fait, les locutions adverbiales de mode véhiculent plusieurs valeurs sémantiques. Il est, selon nous, préférable d'analyser le syntagme prépositionnel ***em massa*** (en masse) comme étant l'expression de l'idée de *en grande quantité ou intensité*, tout en faisant ressortir la valeur quantitative de cette locution.

En nous appuyant sur Guimier qui considère que « Les intensifs sont tout d'abord des adverbes formés sur des adjectifs qui appartiennent déjà, de par leur sens de base, au domaine du quantitatif » (1996 : 29), nous pouvons dire que ***em massa*** (en masse) est une locution adverbiale formée en langue portugaise à partir d'un nom qui comporte le trait sémantique *quantité*.

Sur le plan fonctionnel, les *compléments circonstanciels* sont des *locutions adverbiales nucléaires* (v. *supra*, p. 35). Cette étiquette met en relief le type de relation que l'argument a dans la structure de l'UTC. Autrement dit, les *locutions adverbiales nucléaires* fonctionnent à l'intérieur de la prédication nucléaire puisque leur rapport est établi

directement avec la base de l'UTC. Nous allons voir lorsque nous analyserons d'autres locutions adverbiales, plus loin dans ce chapitre, que certaines ont une relation avec la prédication nucléaire prise comme un tout, prouvant ainsi un autre type de comportement de l'adverbe dans la structure de l'UTC.

Passons maintenant à l'analyse détaillée du comportement de la fonction sémantique des arguments de la règle [ϕ [x_1]]. Les types de fonctions sémantiques relevées dans cet ensemble sont variés. Le tableau 4.11 fait état de toutes ces fonctions sémantiques et présente les pourcentages observés pour chacune.

Tableau 4.11 — Pourcentage d'arguments par fonction sémantique pour la règle [ϕ [x_1]]

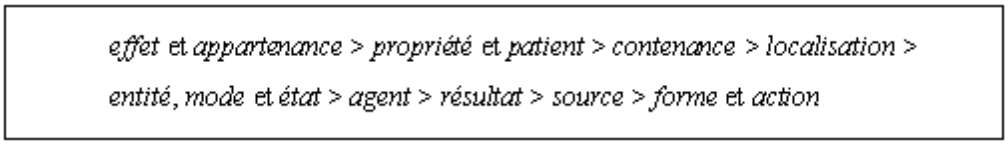
Fonction sémantique de l'argument	**Nombre d'arguments par fonction sémantique**	**%**
action	1	1,00
agent	4	4,00
appartenance	13	13,00
contenance	12	12,00
effet	14	14,00
entité	5	5,00
état	4	4,00
forme	1	1,00
intensité	1	1,00
localisation	6	6,00
mode	10	10,00
patient	12	12,00
propriété	12	12,00
résultat	3	3,00
source	2	2,00
Total	100	100

À la lecture de ce tableau, on observe tout d'abord que le cadre d'application des fonctions sémantiques des arguments est très différent si on le compare à celui de la base (v. tableau 4.7, p. 132). En effet, l'analyse des fonctions sémantiques des arguments et des

bases montre à quel point la typologie sémantique est influencée par le fonctionnement syntaxique des composants d'une UTC. Par exemple, la fonction *entité* est nettement inférieure pour les arguments. Elle ne représente que 5 % alors qu'elle atteignait 41 % dans le cas des bases-sujet. Les arguments se caractérisent principalement par l'emploi des fonctions *effet* (14 %), *appartenance* (13 %), *contenance* (12 %), *propriété* (12 %) et *patient* (12 %).

L'analyse des fonctions sémantiques propres aux UTC, dont la fonction syntaxique est *adjoint de nom* (la plus représentative pour l'ensemble des arguments), fait ressortir la fréquence de la fonction *effet* et *appartenance*. Nous avons illustré ci-dessous cette hiérarchie des fonctions sémantiques, qui s'éloigne sensiblement de celle de la *base-sujet* (v. figure 4.1, p. 132).

Figure 4.2 — Hiérarchie de la fonction sémantique pour les arguments de fonction syntaxique adjoint de nom pour la règle [ϕ [x_1]]



La relation entre ces composants et la position qu'ils occupent dans la structure interne de l'UTC sont des éléments d'analyse essentiels pour une étude comme la nôtre. Ainsi, un même argument peut avoir une fonction sémantique différente selon son entourage terminologique. Prenons, par exemple, l'argument *celular* (cellulaire) dans les UTC *agregado celular* (agrégat cellulaire) et *alongamento celular* (élongation cellulaire). Dans le premier cas, *celular* (cellulaire) a la fonction sémantique *contenance* puisque il s'agit d'un agrégat qui contient des cellules. Dans le second cas, ce même argument a la fonction sémantique *patient* étant donné que *alongamento* (élongation) est un processus qui agit sur la cellule. L'argument *celular* (cellulaire) est alors un patient qui subit l'action produite par le processus d'élongation.

L’argument *somática* (somatique) est un autre exemple que nous avons retenu et qui est illustré dans les UTC reproduites dans le tableau qui suit.

Tableau 4.12 — Fonctions sémantiques de l’argument somática

UTC formées de l’argument *somática*	**Fonction sémantique**
célula ***somática*** (cellule somatique)	appartenance
embrião ***somático*** (embryon somatique)	source
embriogênese ***somática*** (embryogenèse somatique)	patient
hibridação ***somática*** (hybridation somatique)	patient
híbrido ***somático*** (hybride somatique)	source
mutação ***somática*** (mutation somatique)	patient

Dans le cas de *embriogênese* ***somática*** (embryogenèse somatique), *hibridação* ***somática*** (hybridation somatique) et *mutação* ***somática*** (mutation somatique), l’argument est un *patient* puisque la base est un *processus*. Ainsi, *embriogênese (*embryogenèse), *hibridação* (hybridation) et *mutação* (mutation) sont des processus qui agissent sur des cellules somatiques. Même si l’élément *célula* (cellule) n’apparaît pas textuellement dans la formation de ces UTC, il est sous-jacent à l’analyse de celles-ci. En effet, c’est grâce à cet élément qu’on peut affirmer qu’il s’agit d’un *patient*.

Dans le cas de *célula* ***somática*** (cellule somatique), l’argument ***somática*** (somatique) a la fonction *appartenance* compte tenu du fait que la notion exprimée par cette UTC est celle d’une cellule que appartient au soma, au corps.

L'argument ***somático*** a une troisième fonction sémantique dans les UTC *embrião* ***somático*** (embryon somatique) et *híbrido* ***somático*** (hybride somatique). Dans ces deux derniers cas, ***somático*** (somatique) a la fonction de *provenance* ou de *source* parce qu’il comprend la notion de cellule somatique et qu’il constitue la matière qui a donné comme résultat les entités définies par les bases (l’embryon et l'hybride). Ces deux exemples montrent encore une fois, la notion de dépendance qui existe entre les composants des

UTC. L'influence de la notion d'embryon et d'hybride est évidente dans cette analyse. Hybride, par exemple, est défini comme un « **produit** du croisement de deux individus différant par un ou plusieurs caractères héréditaires » (v. Quirion et Bourbeau, s.d.). Par conséquent, les bases *híbrido* (hybride) et *embrião* (embryon) sont des produits ou des résultats de la matière ou de la source appelée cellule somatique. Selon les définitions relevées dans Termium, l'hybride et l'embryon sont tous deux issus des cellules somatiques.

Considérons un dernier exemple, celui de l'argument *floral* (floral) dans les UTC *botão floral* (bourgeon floral), *indução floral* (induction florale) et *iniciação floral* (initiation florale). Le tableau suivant résume les fonctions sémantiques de cet argument dans chaque UTC citée.

Tableau 4.13 — Fonctions sémantiques de l'argument floral

UTC comportant l'argument *floral*	**Fonction sémantique**
botão ***floral*** (bourgeon floral)	*appartenance*
indução ***floral*** (induction florale)	*résultat*
iniciação ***floral*** (initiation florale)	*patient*

Dans ces trois exemples, l'argument change de fonction sémantique selon le rapport qu'il entretient avec la base. Ainsi, *botão floral* (bourgeon floral) est un bourgeon qui appartient à la fleur, *indução floral* (induction florale[41]) est un processus qui résulte en une floraison et *iniciação floral* (initiation florale[42]) est un processus qui agit sur la fleur.

Au niveau pragmatique de l'analyse fonctionnelle, l'argument remplit toujours la fonction dite *focus*. Nous avons défini cette fonction comme appartenant à l'élément de l'UTC qui spécifie le statut informationnel de l'élément *topique* (v. *supra* p. 37). En effet,

[41] Dans Quirion et Bourbeau, induction florale signifie stimulation des phénomènes physiologiques qui président à la floraison.

[42] Dans Termium, initiation florale (syn. mise à fleur) : transformation du méristème qui marque un nouvelle phase dans le développement.

le rôle de l'élément *focus* est de compléter ou de restreindre l'information véhiculée par la base de l'UTC. Étant donné que cette fonction pragmatique est inhérente à la nature propre des arguments, nous allons voir, au cours de notre analyse, qu'elle sera toujours attribuée aux arguments de la structure de l'UTC.

4.2.1.3 Analyse morphosyntaxique

Dans le cadre de l'analyse de la structure syntagmatique des UTC formées selon la règle [ϕ [x_1]], nous nous sommes penchée sur deux types de structures :

- [N [SA]];
- [N [SP]].

L'analyse de ces structures fait ressortir les résultats obtenus pour l'attribution des catégories grammaticales des UTC (v. tableau 4.9, *supra*, p. 135). Ainsi, les bases sont toujours des *noms* et les arguments peuvent être soit un *syntagme adjectival* (SA) soit un *syntagme prépositionnel* (SP). Dans le cas de la règle analysée, il y a une nette prédominance des structures composées d'un *syntagme adjectival* comme le montre le tableau 4.14.

Tableau 4.14 — Pourcentage de SA ou de SP par formule syntaxique pour la règle [ϕ [x_1]]

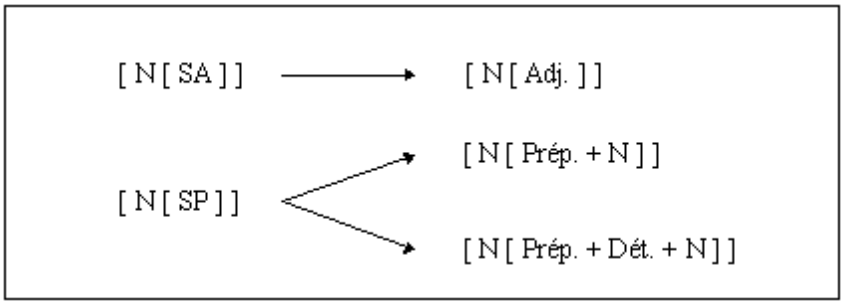
Type de formule syntaxique	**Nombre de formule syntaxique par UTC**	**%**
[N [SA]]	81	81,00
[N [SP]]	19	19,00
Total	100	100

La prédominance des structures qui prennent la forme de [N [SA]] a déjà été notée par d'autres auteurs. C'est ainsi que Guilbert affirme que « [...] l'adjectif apparaît, en

effet, comme l'instrument le plus apte à fournir les réalisations particulières et différenciées d'unités lexicales préexistantes. Son emploi dans cette fonction constitue une forme du processus de dénomination qui consiste à désigner une chose par son trait le plus significatif » (1965 : 270).

À lumière de ces formules syntaxiques, nous avons analysé le comportement morphosyntaxique des UTC. Cette analyse a été synthétisée en formules morphologiques. Nous présentons ci-dessous les formules morphologiques relevées par rapport aux structures syntaxiques examinées.

Figure 4.3 — Formules syntaxiques et représentation en formules morphologiques de la règle [ϕ [x_1]]



Les UTC qui suivent renvoient à la formule [N [Adj.]] :

- *bactéria anaeróbica* (bactérie anaérobie);
- *broto apical* (bourgeon apical);
- *célula somática* (cellule somatique);
- *embrião somático* (embryon somatique);
- *hibridação somática* (hybridation somatique);
- *mapa genético* (carte génétique);
- *propagação clonal* (propagation clonale);
- *tecido diferenciado* (tissu différencié);
- *vetor genético* (vecteur génétique).

Toutefois, cette formule englobe également les UTC dont l'argument est formé d'un préfixe de négation suivi d'un adjectif. Ce préfixe de négation est représenté dans notre corpus par le formant *não* (non). Les recherches menées aux fins de l'analyse de la formation terminologique ont démontré que ce préfixe de négation est un type de formant présent dans divers domaines de connaissances. En effet, le recours à ce préfixe n'est pas un phénomène récent. Guilbert signale qu'en français « [...] la négation par la préfixation de *non* avait commencé à se développer au XVIII^{e} siècle. [...] ce procédé apparaît dans le langage juridique et surtout politique au moment de la Révolution » (*id* : 310).

D'autres auteurs ont également fait mention de ce préfixe dans la formation de termes. Kocourek, par exemple, classe le préfixe *non-* parmi les morphèmes qui entrent dans la formation terminologique. L'auteur place ce formant parmi les « [...] dix-huit particules antérieures qui se situent entre l'affixation et la composition, et qui sont traditionnellement rangées parmi les préfixes » (1991 : 113).

Nous croyons que le préfixe portugais *não* a pour fonction de marquer les rapports sémantiques entre les arguments et la base d'une UTC. Ainsi, à l'intérieur de l'argument de l'UTC, nous analysons l'argument formé par la séquence *préfixe plus adjectif* comme un processus de dérivation qui se forme dans la composition terminologique de l'UTC. Ainsi, lié morphologiquement à l'adjectif, ce préfixe a pour rôle sémantique de nier un trait significatif de la base de l'UTC. Ce processus de dénomination a pour but de préciser le concept de la base au moyen de la négation, tout en marquant le rapport sémantique entre l'argument et la base.

C'est dans ce sens que Sager mentionne que

> Functionally, derivation and compounding serve the purpose of closer determination of a concept - narrowing its intention - while at the same time showing the relationship that exists between the new concept and its origin. In this way determination can make transparent a particular dimension of the

> conceptual structure. This is both its strength and its weakness: by expressing one relationship, it is silent on others which conceptually may be considered more significant (1991 : 73).

Dans son explication, Sager cite la formation terminologique au moyen des préfixes de négation comme une façon de créer des « opposite terms » (v. *id*. : 76).

Pour sa part, Guilbert relève le caractère de clarté de ce type de formation qui utilise la négation détachée de l'adjectif et précise également que cette construction « permet des oppositions significatives sans équivoque » (1965 : 310).

Nous citons ci-dessous quelques exemples trouvés dans notre corpus :

- *bactéria* ***não*** *virulenta* (bactérie non virulente);
- *célula* ***não*** *diferenciada* (cellule indifférenciée);
- *cromossomo* ***não****-homólogo* (chromosome non-homologue);
- *cromossomo* ***não****-pareado* (chromosome non-apparié);
- *DNA* ***não*** *repetido* (DNA non répété);
- *DNA* ***não*** *selecionável* (ADN non sélectionné);
- *genótipo* ***não*** *indutivo* (génotype non inductif);
- *método* ***não*** *enzimático* (méthode non enzymatique).

De cette liste d'exemples, ce qui retient notre attention c'est l'utilisation instable du trait d'union. L'indécision quant à l'emploi du trait d'union peut se justifier par l'influence qu'a la langue anglaise sur le portugais. En effet, le portugais s'appuie parfois sur l'anglais pour constituer sa terminologie dans le domaine de la biotechnologie des végétaux. En anglais, les préfixes « [...] are more like compound elements in that they are often joined to the base by a hyphen, especially if they are 'strong' prefixes which normally receive full stress (non-, pseudo-, ultra-, anti-, etc.) » (Quik et *al*. 1972 : 980).

Nous voulons également mettre en évidence que pour la plupart des UTC formées d'un préfixe de négation, la forme positive correspondante existe également. Cette façon de faire propre à certains domaines spécifiques a été relevée à quelques reprises dans notre corpus. Tels sont les cas des UTC *método* ***não*** *enzimático* (méthode non enzymatique) et *tecido* ***não*** *meristemático* (tissu non méristématique) qui ont, respectivement, les formes opposées suivantes : *método enzimático* (méthode enzymatique) et *tecido meristemático* (tissu méristématique).

Si l'ajout d'un adjectif a pour fonction de préciser le concept de la base de l'UTC en faisant ressortir un trait significatif, l'emploi du préfixe de négation suivi d'un adjectif a pour rôle de nier ce trait significatif de la base, ce qui est également une façon de préciser un concept donné.

Passons maintenant à l'analyse de la formule syntaxique [N [SP]]. Comme nous l'avons vu dans la figure 4.3, celle-ci est représentée soit par la forme [N [Prép. + N]] soit par [N [Prép. + Dét. + N]].

Les UTC formées selon la structure du type [N [Prép. + Dét. + N]] ont été retenues puisqu'elles figurent dans des dictionnaires. Nous avons relevé, dans notre corpus, trois UTC caractérisées par le modèle [N [Prép. + Dét. + N]] :

- *ápice do*[43] *broto*[44] (extrémité de la tige);
- *axila da*[45] *folha*[46] (aisselle de la feuille);
- *base da*[47] *folha*[48] (base de la feuille).

---

[43] En portugais, l'élément *do* est une contraction de la préposition *de* et de l'article *o*.

[44] Le terme *ápice do broto* (extrémité de la tige) est l'entrée du glossaire présenté à la fin du livre utilisé comme corpus pour la collecte des données. Même l'équivalent en français (trouvé dans Termium) garde la préposition suivie de l'article.

[45] En portugais, l'élément *da* est une contraction de la préposition *de* et de l'article *a*.

[46] Nous avons trouvé l'équivalent en français dans la banque de terminologie Termium qui conservait le même patron N + Prép. + Dét. + N que *axila da folha* (aisselle de la feuille) .

[47] Idem note 9.

[48] Le terme *base da folha* (base de la feuille) a été attesté dans le *Novo dicionário da língua portuguesa* comme appartenant au domaine *Botanique*.

Ces exemples constituent certainement des exceptions à la règle, toutefois ils doivent être pris en considération puisqu'ils constituent des unités complexes déjà lexicalisées dans le domaine. Autrement dit, ces exemples viennent prouver que l'absence d'article devant la prédication n'est pas toujours un élément qui caractérise le figement des UTC.

Dans notre corpus, nous avons détecté des variantes d'une même UTC qui prennent soit la forme syntaxique [N [SP]] soit [N [SA]]. Autrement dit, les arguments formés de *Prép. + N* prennent parfois la forme d'adjectifs et vice versa, tout en gardant la même valeur sémantique. Ce phénomène a été étudié par Faulstich qui les classe comme des variantes terminologiques linguistiques du type syntaxique. Selon l'auteure, « la variation est effectuée par la substitution d'une partie de l'élément terminologique par une structure semblable trouvée dans le lexique de la langue »[49] (v. s. d. a : 3). Quelques exemples de ce type de variantes syntaxiques sont présentés dans le tableau ci-dessous.

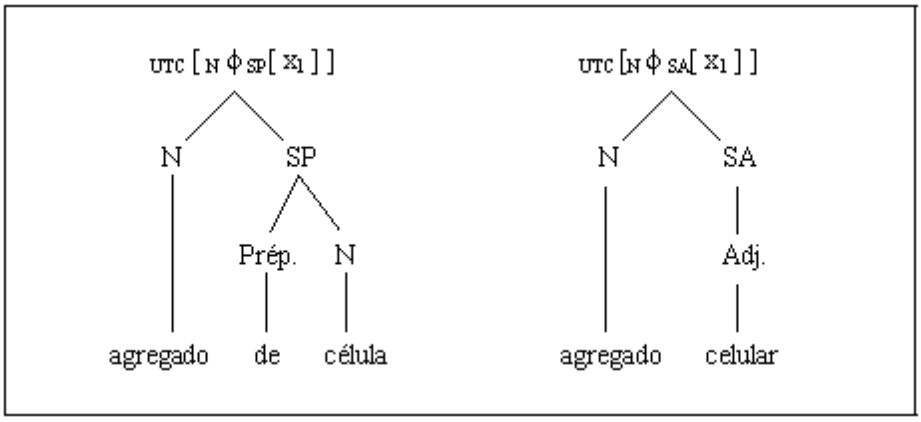
Tableau 4.15 — Changement de structures de l'argument pour un même concept

[N [SP]]	**[N [SA]]**
agregado de células (agrégat de cellules)	*agregado celular* (agrégat cellulaire)
axila da folha (aisselle de la feuille)	*axila foliar* (aisselle foliaire)

Ainsi, pour une même règle de formation et un même concept, nous pouvons avoir différents arbres syntagmatiques. Prenons les variantes *agregado de células* (agrégat de cellules) et *agregado celular* (agrégat cellulaire) de la figure 4.4.

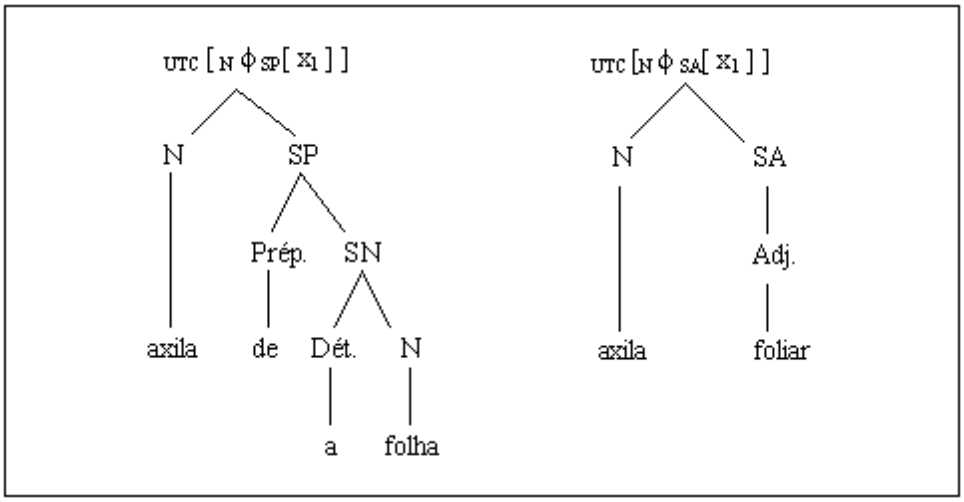
[49] Traduction personnelle du texte original écrit en portugais.

Figure 4.4 — Arbres syntagmatiques d'agregado de célula/agregado celular



La représentation en arbre syntagmatique des UTC *axila da folha* (aisselle de la feuille) et *axila foliar* (aisselle foliaire) se distingue un peu de l'exemple précédent. Voyons le résultat présenté dans la figure 4.5.

Figure 4.5 — Arbres syntagmatiques d'axila da folha/axila foliar



4.2.1.4 Relation conceptuelle entre la base et les arguments

La fonctionnalité de la langue, en terminologie, est un phénomène qui doit être transparent pour atteindre la précision nécessaire dans les rapports de communication entre spécialistes. Dans la perpective de la grammaire fonctionnelle, les UTC sont étudiées sur une base pragmatique à l'intérieur de laquelle la sémantique et la syntaxe sont analysées. Ainsi, les relations conceptuelles reflètent cette base pragmatique de sorte que la fonction sémantique de chaque élément de l'UTC est perçue de différentes manières selon la relation conceptuelle associée par l'UTC prise comme un tout.

L'analyse des relations conceptuelles des UTC formées selon la règle [ϕ [x_1]] a permis de dégager 13 types de relations : *agit comme un, agit de manière à, agit sur, appartient à, contient, est du type, est en état, produit par, produit un effet, produit un résultat, résulte de, se situe dans, subit l'action.*

Nous avons également noté que chaque relation conceptuelle comporte un sous-ensemble de fonctions sémantiques représentées par les éléments de l'UTC et que ces sous-ensembles sont plus au moins constants. Nous avons reproduit dans le tableau suivant, les relations conceptuelles ainsi que les fonctions sémantiques correspondantes.

Tableau 4.16 — Relations conceptuelles et fonctions sémantiques pour la règle [ϕ [x_1]]

Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**	
	base	**argument**
agit comme un	*agent*	*propriété*
agit de manière à	*processus* *processus*	*mode* *intensité*
agit sur	*processus*	*patient*
appartient à	*agent* *entité* *processus*	*appartenance* *appartenance* *appartenance*
contient	*groupement* *entité*	*contenance* *contenance*
est du type	*entité* *entité*	*propriété* *forme*
est en état	*entité* *localisation*	*état* *état*
produit par	*processus* *entité*	*agent* *mode*
produit un effet	*processus* *agent*	*effet* *effet*
produit un résultat	*processus*	*résultat*
résulte de	*entité* *effet*	*source* *propriété*
se situe dans	*processus* *localisation* *entité*	*localisation* *entité* *localisation*
subit l'action	*localisation*	*action*

Une analyse au niveau sémantico-conceptuel n'est pas simple. Le tableau ci-dessus comporte des zones nébuleuses malgré les efforts que nous avons faits pour suivre des paramètres établis avec précision.

Certains cas ne nous ont causé aucun problème d'analyse, notamment la relation *agit comme un*. Cette relation résulte d'un rapport entre une base dont la fonction est *agent*

et un argument dont la fonction est *propriété*[50]. L'UTC *célula hospedeira* (cellule hôte) illustre cette relation puisque la cellule, qui est l'agent dans cette UTC, *agit comme un* hôte.

Il faut signaler que la relation *agit comme un* se distingue de la relation *agit de manière à* qui découle des rapports entre *processus + mode* et *processus + intensité*. C'est la façon d'effectuer le processus qui est mis en cause dans cet exemple. Les UTC *autofecundação controlada* (autofécondation contrôlée) et *cultura em batelada* (culture en batch) sont des exemples de *processus + mode*. Dans le cas de *propagação em massa* (propagation en masse), l'argument *em massa* (en masse) a la fonction sémantique *intensité*. Pourtant, certains auraient pu l'interpréter comme un *mode*. Néanmoins, c'est le trait sémantique *quantité* qui le distingue des autres et qui lui confère la fonction *intensité*.

La relation *agit sur* est un autre exemple clair à visualiser. Cette relation met en rapport les fonctions *processus* et *patient* présentes dans des UTC comme

- *alongamento celular* (élongation cellulaire);
- *alteração cromossômica* (altération chromosomique);
- *análise genômica* (analyse génomique).

Dans le cas de la relation *contient*, elle peut marquer un rapport entre un *groupement* ou une *entité* en relation à la fonction de *contenance* exprimée par l'argument de l'UTC. Ainsi, l'UTC *agregado celular* (agrégat cellulaire) met en relation les fonctions *groupement + contenance* tandis que dans *casa de vegetação* (serre) et *bactéria não virulenta* (bactérie non virulente), il y a un rapport *entité + contenance*. En effet, pour *bactéria não virulenta* (bactérie non virulente), le rapport de *contenance* est négatif. Autrement dit, il s'agit de la non-contenance du virus dans la bactérie.

[50] La fonction sémantique *propriété* comprend la notion d' « ensemble de constances, de caractères [surtout caractère de fonction], de réactions d'une substance; manière dont elle se comporte suivant les conditions dans lesquelles elle est placée » (v. Le Nouveau Petit Robert, 1995).

Pour ce qui est de la relation *se situe dans*, elle peut exprimer trois types de rapports, à savoir :

- *processus* + *localisation*
  ex.: *ação endógena*[51] (action endogène).

- *localisation* + *entité*
  ex : ápice do broto (extrémité de la tige);
  axila da folha (aisselle de la feuille);
  axila foliar (aisselle foliaire).

- *entité* + *localisation*
  ex : gemas adventícias (gemme adventice);
  genoma nuclear (génome nucléaire);
  meristema apical (méristème apical).

Il est intéressant de voir, encore une fois, comment un même argument change de fonction sémantique selon le rapport conceptuel exprimé par l'UTC. Par exemple, l'argument *nuclear* (nucléaire) a la fonction sémantique de *localisation* dans *genoma nuclear* (génome nucléaire), mais change pour la fonction *patient* dans l'UTC *alteração nuclear* (altération nucléaire) et exprime ainsi un autre type de rapport conceptuel, à savoir le rapport *agit sur*. La fonction sémantique de la base de ces deux UTC a d'ailleurs une grande influence sur le type de rapport entre leurs composants.

Pour ce qui est du rapport *est en état*, c'est la séquence *entité* + *état* qui compte le plus d'exemples. Nous n'avons relevé qu'une seule UTC de la séquence *localisation* + *état*, soit *sítio ativo* (site actif). Voici quelques UTC qui illustrent le type *entité* + *état* :

[51] L'action se situe à l'intérieur d'un organisme.

- *antera madura* (anthère mature);
- *genótipo transitório* (génotype transitoire);
- *tecido embrionário* (tissu embryonnaire).

La relation *est du type* ne se vérifie pas facilement. À première vue, toutes les relations semblent illustrer ce rapport. Afin de ne pas fausser les résultats, nous avons défini des critères pour reconnaître la relation *est du type*. Ainsi, pour être considérée comme une relation *est du type*, il faut que la fonction sémantique de la base soit *entité* et celle de l'argument *propriété*. La seule exception a été l'UTC *híbrido anão* (hybride nain) qui marque le rapport entre les fonctions *entité* et *forme*.

L'UTC *genótipo superior* (génotype supérieur) illustre la relation *est du type* qui met en rapport les fonctions sémantiques *entité* et *propriété*. La fonction sémantique *propriété* peut également être exprimée par une négation comme le montrent les exemples qui suivent :

- *célula não diferenciada* (cellule indifférenciée);
- *cromossomo não-homólogo* (chromosome non-homologue);
- *cromossomo não-pareado* (chromosome non-apparié);
- *DNA não repetido* (DNA non répété);
- *tecido não meristemático* (tissu non méristématique).

En ce qui concerne la relation *appartient à*, elle est exprimée par des arguments du type *appartenance* qui ont un rapport avec des bases dont la fonction est *agent*, *entité* ou *processus*. La fonction d'*appartenance* est définie comme une situation qui lie un agent, une entité ou un processus à une collectivité ou à un ensemble. La plupart des arguments qui ont reçu l'étiquette *appartenance* font référence à un domaine ou à une science, ici interprété comme un ensemble de connaissances. Toutefois, cette analyse dépend du rapport entre l'argument et la base de l'UTC. Ainsi, l'argument *genética* (génétique) a reçu dans certains cas la fonction *appartenance* et dans d'autres, la fonction *contenance* ou *patient*

selon le rapport conceptuel exprimé par l'UTC. Lorsque la fonction marque la *contenance* comme dans *mapa genético* (carte génétique), le rapport est *contient*. Toutefois, il y a des UTC dont l'argument comprend plus que la notion de « gène ». Cette notion plus élargie porte parfois sur l'information génétique ou encore sur la science de la génétique. Voici des exemples de ce type d'argument qui représente la fonction *appartenance* :

- *anormalidade genética* (anomalie génétique);
- *base genética* (base génétique);
- *estabilidade genética* (stabilité génétique);
- *ganho genético* (gain génétique);
- *marcador genético* (marqueur génétique);
- *variabilidade genética* (variabilité génétique).

D'autres arguments expriment la fonction d'*appartenance* dans un rapport du type *appartient à*. Ces arguments, qui font parfois référence à un domaine de connaissances, peut être combiné avec des bases dont la fonction est *agent*, *entité* ou *processus*. Ainsi les exemples suivants :

- *agent* + *appartenance*
  ex. : *agente biológico* (agent biologique).

- *processus* + *appartenance*
  ex. : *aplicação biotecnológica* (application biotechnologique).

- *entité* + *appartenance*
  ex. : *botão floral* (bourgeon floral);
  célula somática (cellule somatique);
  método de propagação (méthode de propagation).

Quant à la relation *résulte de*, elle reflète, comme le montre le tableau 4.16, un rapport entre une *entité* et une *source* ou entre un *effet* et une *propriété*. Voici quelques exemples :

- *entité + source*

  ex. : *embrião somático* (embryon somatique) ⇒ l'embryon **résulte de** cellules somatiques

  híbrido somático (hybride somatique) ⇒ l'hybride **résulte des** cellules somatiques

- *effet + propriété*

  ex. : efeito inibitório (effet inhibiteur) ⇒ l'effet **résulte de** l'inhibition.

Analysons maintenant le rapport conceptuel *produit un résultat*. Ce rapport est toujours établi entre une base qui a la fonction *processus* et un argument dont la fonction est *résultat*. Trois exemples illustrent ce rapport :

- *cultura de anteras* (culture d'anthères);
- *indução floral* (induction florale);
- *propagação clonal* (propagation clonale).

Ce rapport ne doit pas être confondu avec *produit un effet*. Dans le premier, l'argument est une chose (ex : une anthère, une fleur, un clone), tandis que dans *produit un effet* l'argument exprime une notion dynamique.

Le rapport *produit un effet* se réalise par la combinaison d'une base exprimant un *processus* ou un *agent* et l'argument qui a la fonction sémantique *effet*. Les exemples ci-dessous illustrent ce rapport.

- *processus* + *effet*

  ex. : *ação desagregadora* (action de désagrégation);

  mecanismo de regulação (mécanisme de régulation).

- *agent* + *effet*

  ex. : *agente infectivo* (agent infectieux);

  agente patogênico (agent pathogène);

  genótipo não indutivo (génotype non inductif);

  vetor de clonagem (vecteur de clonage).

En ce qui concerne le rapport *produit par*, il est exprimé par la relation entre *processus* + *agent* et par *entité* + *mode*. Nous présentons ci-dessous les exemples qui illustrent ce type de rapport.

- *processus* + *agent*

  ex. : *ação enzimática* (action enzymatique);

  reinfecção viral (réinfection virale);

  reinfecção virótica (réinfection virale).

- *entité* + *mode*

  ex. : *auxina sintética* (auxine de synthèse);

  híbrido por complementação (hybride par complémentation);

  célula em suspensão[52] (cellule en suspension).

Finalement, le rapport *subit l'action* a été attribué à l'UTC *sítio de restrição* (site de restriction) qui comprend une relation entre une base dont la fonction est *localisation* et l'argument *restrição* (restriction) qui a la fonction *action*. Nous avons défini la fonction *action* pour cet argument puisque *restrição* (restriction) fait appel à l'action produite par

[52] La cellule est produite par la méthode de culture en suspension

l'enzyme de restriction. Cette analyse est fondée sur la définition de l'UTC *sítio de restrição* (site de restriction) : « séquence d'ADN double-brin spécifiquement **reconnue et clivée par une enzyme de restriction** donnée » (Quirion et Bourbeau, s.d.).

4.2.1.5 Synthèse de la règle [ϕ [x_1]]

Au cours de cette section, il été question des UTC formées selon la règle [ϕ [x_1]]. Étant donné la fréquence élevée de ce type de formation, ce sous-ensemble de notre corpus a occupé une large place dans nos analyses. Ce qui retient d'abord notre attention c'est l'importante majorité des *bases statiques* qui figurent dans notre corpus. Le recours à ce type de bases reflète le comportement du discours terminologique qui privilégie l'emploi de la nominalisation et non des structures verbales. Quant aux *bases actives*, nous n'en avons relevé que quelques exemples, mais il faut mentionner que celles-ci ont, tout comme les *bases statiques*, un comportement discursif nominatif. En effet, ce qui distingue ces deux types de bases c'est le fait que les *bases actives* sont des noms dérivés de verbes, tandis que les *bases statiques* sont des noms dès le départ.

L'analyse de nos données a également fait ressortir des tendances différentes quant à l'emploi des fonctions sémantiques relatives aux bases et aux arguments. Nous remarquons notamment que pour les bases, la fonction *entité* a une fréquence d'utilisation de 41 %, tandis que pour les arguments cette fonction ne représente que 5 %. Comme nous l'avons vu à plusieurs reprises dans cette section, la fonction sémantique des composants varie aussi selon les rapports conceptuels internes de l'UTC.

En ce qui concerne les patrons syntaxiques, c'est le patron [N [SA]] qui est le plus utilisé pour la formation des UTC puisque les catégories grammaticales *nom* pour les bases et *adjectif* pour les arguments prédominent. Fait intéressant à noter, parmi les arguments composés d'un *adjectif*, nous avons relevé plusieurs exemples qui comportent le préfixe de

négation *não* qui sert à marquer les rapports sémantiques entre les arguments et la base d'une UTC.

Quant au patron [N [SP]], il regroupe, au-delà des formules [N [Prép. + N]], des UTC lexicalisées du type [N [Prép. + Dét. + N]]. Celles-ci viennent d'ailleurs confirmer la possibilité d'avoir des structures figées qui comportent des articles à l'intérieur des arguments. L'analyse du patron [N [SP]] a également démontré la valeur de la préposition comme élément représentatif des rapports entre les composants de l'UTC.

En ce qui a trait aux fonctions syntaxiques des UTC, on en relève quatre : *sujet* pour la base, *adjoint de nom*, *complément du nom* et *complément circonstanciel* pour les arguments. Les *compléments circonstanciels* se distinguent des trois autres types puisqu'ils sont des arguments qui ont un rapport direct avec la base et qui ont un rôle de locution adverbiale de mode.

Finalement, nous avons regroupé dans le tableau 4.16 les treize relations conceptuelles relatives à la règle [ϕ [x_1]]. Pour chacune de ces relations, nous avons réussi à systématiser les fonctions sémantiques de la base et des arguments qui s'y rattachent. Ce tableau démontre donc la cohérence et la constance des fonctions sémantiques propres aux relations conceptuelles exprimées par les UTC de la règle [ϕ [x_1]].

4.2.2 [[ϕ [x_1]] [y_1]]

Dans la présente section, nous allons analyser les principaux aspects de la règle [[ϕ [x_1]] [y_1]]. Malgré le faible nombre d'exemples recensés dans notre corpus, cette règle est importante puisqu'elle révèle un mode de formation propre à la biotechnologie des végétaux, mais qui peut également se retrouver dans d'autres domaines. Nous croyons d'ailleurs qu'il suffit de repérer un seul exemple en contexte pour qu'une règle soit prise en

considération. De plus, nous verrons plus loin que cette règle sert de point de départ à un autre modèle.

La règle de formation [[ϕ [x_1]] [y_1]] représente les UTC du type *base* + *argument* + *satellite*. Elle est composée d'une *prédication nucléaire* et d'une *prédication étendue*. L'ensemble *base* + *argument* constitue la *prédication nucléaire*. L'ajout du satellite à cet ensemble forme la *prédication étendue*.

En ce qui concerne les types de bases, les exemples recueillis offrent, contrairement à la règle [ϕ [x_1]], un portrait homogène quant à l'usage des bases actives et statiques. Au niveau de l'emploi des affixes, il a été intéressant d'observer la fréquence élevée d'utilisation du suffixe *-ção*. Cette constatation va de pair avec le fait que les bases régies par cette règle expriment une fonction sémantique *processus*, ce qui est un des rôles du suffixe en question. Voici quelques exemples d'emploi de ce suffixe.

- *aclimata**ção** de plantas no solo* (acclimatation des plantes au sol);
- *recombina**ção** homóloga in vivo* (recombinaison homologue in vivo);
- *indu**ção** floral in vitro* (induction florale in vitro).

Quant aux arguments, ceux-ci ont un emploi presque équivalent pour ce qui est des catégories grammaticales. Ils sont soit des a*djectifs*, soit des *syntagmes prépositionnels*. Cette caractéristique diffère de la règle antérieure où les arguments adjectivaux étaient plus nombreux. Au niveau syntaxique, ils sont des *adjoints de nom* ou des *compléments du nom*. Pour ce qui est des fonctions sémantiques, les arguments jouent le rôle d'*appartenance*, de *patient*, de *propriété* ou de *résultat*.

Étant donné que l'apport significatif de cette règle est lié à l'ajout du satellite à la structure d'une UTC, nous allons nous concentrer sur cette nouvelle caractéristique dans les pages qui suivent.

4.2.2.1 Interprétation fonctionnelle de la prédication étendue : le satellite

Les satellites sont des éléments qui fournissent plus d'informations à l'ensemble *base + argument*. Ils ont une relation avec la prédication nucléaire prise comme un tout, formant ainsi la prédication étendue. On pourrait les appeler des *modificateurs externes* par rapport aux arguments qui sont considérés comme des *modificateurs internes*[53]. Au niveau virtuel, les satellites ont la possibilité de pouvoir se séparer du reste de l'UTC. C'est comme si on pouvait insérer une virgule entre le dernier argument et le satellite.

Le fait que les satellites aient un rapport avec la prédication nucléaire prise comme un tout est primordial pour les distinguer des arguments. Toutefois, ce n'est pas une tâche facile comme nous l'avons expliqué en détail au premier chapitre (v. *supra*, p. 15).

Dans le cas de la règle [[ϕ [x_1]] [y_1]], les satellites sont soit des *syntagmes prépositionnels* soit des *locutions adverbiales*. Le seul syntagme prépositionnel relevé pour ce qui est des satellites est ***no solo*** (au sol) qui apparaît dans l'UTC *aclimatação de plantas* ***no solo*** (acclimatation des plantes au sol).

Nous avons réservé la catégorie *locution adverbiale* aux expressions latines *in vivo* et *in vitro*. Nous avons considéré ces satellites comme des *locutions adverbiales* pour deux raisons. D'une part, parce que les dictionnaires les classent dans cette catégorie. D'autre part, parce qu'il s'agit d'expressions latines figées dans leur structure et qu'elles ne peuvent pas être analysées comme une séquence *Prép. + N*. Chacune de ces expressions forme un bloc unique qui se distingue du reste de l'UTC. D'ailleurs, dans les textes, elles sont presque toujours en italique ou entre guillemets et, dans les dictionnaires, elles ont leur propre entrée classée selon la première lettre de l'expression. De plus, leur figement est tel qu'elles passent d'une langue à l'autre en conservant leur forme lexicalisée d'origine.

---

[53] Cette nomenclature a été proposée par Perini qui a analysé, pour la langue générale, les fonctions syntaxiques des syntagmes nominaux en portugais (v. Perini, 1996 : 101-106).

Pour ce qui est des fonctions syntaxiques, les satellites sont considérés comme des *compléments circonstanciels*. Il importe de préciser que, pour la règle [ϕ [x_1]], les *compléments circonstanciels* étaient réservés aux arguments appartenant à la catégorie grammaticale syntagme prépositionnel de nature adverbiale (v. *supra*, p. 138).

Dans le cadre de l'analyse fonctionnaliste, les satellites sont des *compléments circonstanciels* qui jouent le rôle d'adverbes ou de *locutions adverbiales étendues* (v. *supra*, p. 34). Nous les appelons ainsi en raison du type de relation que les satellites ont dans la structure de l'UTC. Contrairement aux arguments analysés pour la règle [ϕ [x_1]], qui ont un rapport **dans** la prédication nucléaire et qui sont classés comme des *locutions adverbiales nucléaires* (v. *supra*, p. 35), les satellites de la règle [[ϕ [x_1]] [y_1]] établissent leur rapport **avec** la prédication nucléaire à partir de la prédication étendue. Nous présentons, dans la figure qui suit, deux exemples. Le premier analyse l'UTC *cultura* ***em suspensão*** (culture en suspension) qui est du type [ϕ [x_1]] et qui comporte une *locution adverbiale nucléaire*. Le second illustre l'UTC *recombinação homóloga* ***in vivo*** (recombinaison homologue in vivo) qui suit la règle [[ϕ [x_1]] [y_1]] et qui est composée d'une *locution adverbiale étendue*.

Figure 4.6 — Locution adverbiale nucléaire dans la structure [ϕ [x_1]]

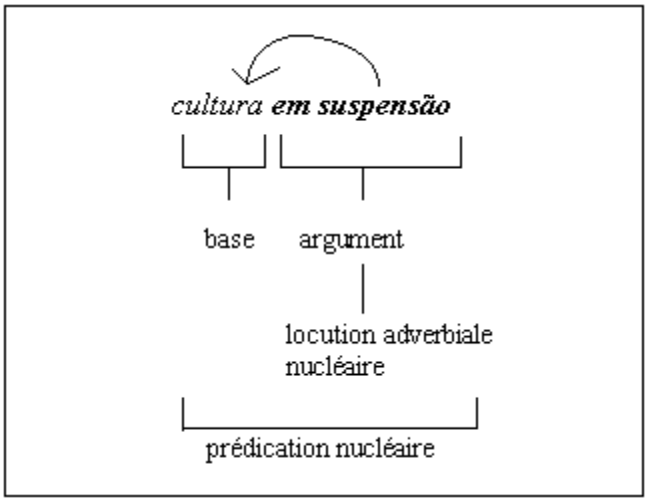
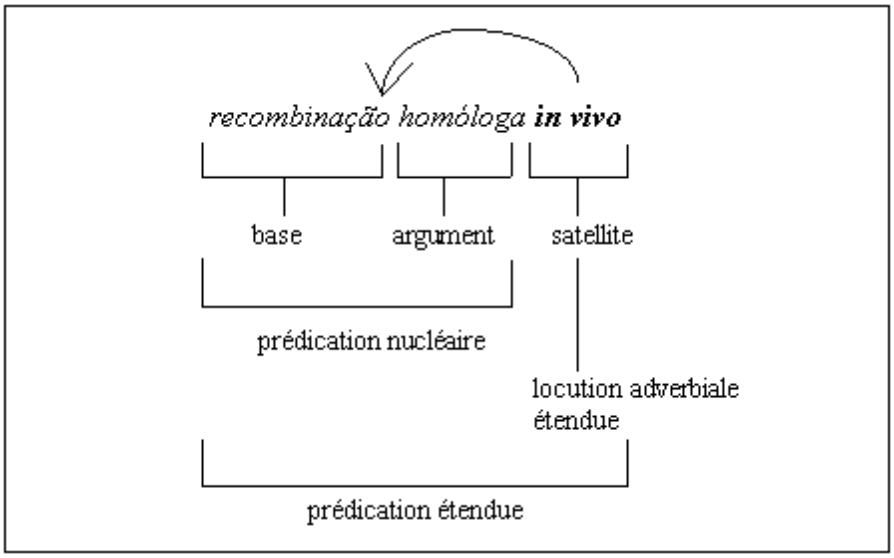


Figure 4.7 — Locution adverbiale étendue dans la structure [[ϕ [x_1]] [y_1]]



Pour ce qui est de la fonction pragmatique, les satellites sont considérés comme des *focus* puisque leur fonction est de spécifier les informations fournies par la prédication nucléaire.

Enfin, les satellites représentent deux types de fonctions sémantiques, à savoir : *état* et *localisation* qui sont illustrées dans le tableau 4.17.

Tableau 4.17 — Fonctions sémantiques des satellites de la règle [[ϕ [x_1]] [y_1]]

Fonction sémantique des satellites	**UTC**
état	*recombinação homóloga* ***in vivo*** (recombinaison homologue in vivo)
localisation	*propagação clonal* ***in vitro*** (propagation clonale in vitro)

Dans le premier exemple, la recombinaison homologue se passe entre gènes clonés vivants (état), tandis que dans le second exemple la propagation clonale est produite en laboratoire (localisation).

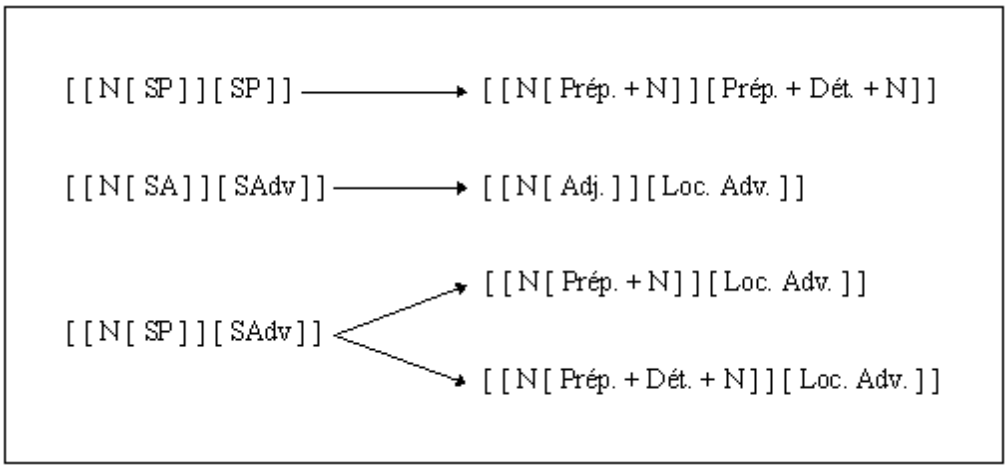
4.2.2.2 Analyse morphosyntaxique

Les UTC de la règle [[ϕ [x_1]] [y_1]] peuvent se présenter sous trois types de structures syntaxiques, à savoir :

- [[N [SP]] [SP]]
- [[N [SA]] [SAdv]]
- [[N [SP]] [SAdv]]

Ces formules syntaxiques peuvent parfois avoir plus d'une forme morphologique comme le montre la figure 4.8.

Figure 4.8 — Formules syntaxiques et représentation en formules morphologiques pour la règle [[ϕ [x_1]] [y_1]]



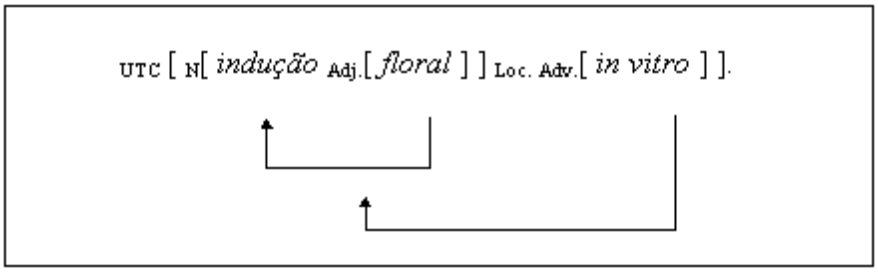
Nous avons regroupé dans le tableau 4.18 des exemples d'UTC pour chacune de ces formules morphologiques.

Tableau 4.18 — Exemples d'UTC par formule morphologique pour la règle [[ϕ [x_1]] [y_1]]

Formule morphologique	**UTC**
[[N [Prép. + N]] [Prép. + Dét. + N]]	*aclimatação de plantas no solo* (acclimatation des plantes au sol)
[[N [Adj.]] [Loc. Adv.]]	*indução floral in vitro* (induction florale in vitro) *propagação clonal in vitro* (propagation clonale in vitro)
[[N [Prép. + N]] [Loc. Adv.]]	*cultura de anteras in vitro* (culture d'anthères in vitro)
[[N [Prép. + Dét. + N]] [Loc. Adv.]]	*tradução do mRNA in vitro* (traduction de l'ARN-m in vitro)

Le recours aux crochets permet d'illustrer, d'une manière claire, les relations établies entre les éléments de l'UTC. Autrement dit, cette façon de faire aide à mieux cerner les ensembles et les sous-ensembles de composants, ce qui facilite la visualisation des rapports. Prenons par exemple l'UTC *indução floral in vitro* (induction florale in vitro). Cette UTC est structurée morphologiquement comme [[N [Adj.]] [Loc. Adv.]]. Si nous substituons les catégories grammaticales par les composants de l'UTC en question, nous allons avoir une représentation du type [[*indução* [*floral*]] [*in vitro*]]. Comme nous pouvons le voir dans la figure 4.9, les ensembles cernés pas les crochets traduisent la relation de *floral* (florale) avec *indução* (induction) formant l'ensemble *indução floral* (induction florale) ainsi que le rapport de *in vitro* établi avec tout le segment qui précède, c'est-à-dire avec *indução floral* (induction florale) pris comme un tout.

Figure 4.9 — Rapports entre les composants de l'UTC pour la règle [[ϕ [x_1]] [y_1]]



D'ailleurs, c'est justement le type de rapport qui permet de savoir si un composant est une base, un argument ou un satellite. Dans l'exemple illustré ci-dessus *indução* (induction) est la base, *floral* (florale) est l'argument et *in vitro* est le satellite.

4.2.2.3 Relation conceptuelle entre la base, l'argument et le satellite

Les rapports que nous venons de voir sont des réalisations morphosyntaxiques des rapports conceptuels établis entre les composants de l'UTC. Ces rapports résultent des relations qui existent entre les fonctions sémantiques des composants. Ces mécanismes ont pour but de représenter une notion donnée.

L'analyse des relations conceptuelles d'une terminologie donnée peut être de diverses natures selon les besoins de l'étude. Sager cite trois types de relations qui sont les plus souvent employées en terminologie : *generic relationships*, *partitive relationships* et *polivalent relationships* (v. 1990 : 30-34). Cependant, l'auteur mentionne que cette analyse n'est pas toujours suffisante. Selon lui,

> Concepts are often seen as being inter-related in a complex manner which cannot be conveniently captured by straightforward generic and partitive structures. [...] The relation between two concepts is bound by the conceptual class of each. For example, relationships of product or material can only exist between material entities; in this way a pattern emerges which shows

> restrictions on the nature of the relationships between concepts by virtue of their categories (*id*. : 34-35)

Dans le cadre de la règle [[ϕ [x_1]] [y_1]], nous avons trouvé trois types de relations conceptuelles. Le tableau ci-dessous les résume et présente les fonctions sémantiques de chacune d'elles.

Tableau 4.19 — Relations conceptuelles et fonctions sémantiques pour la règle [[ϕ [x_1]] [y_1]]

Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**		
	base	**argument**	**satellite**
agit sur / se situe dans	*processus* *agent*	*patient* *patient*	*localisation* *localisation*
est du type / *est en état*	*processus*	*propriété*	*état*
produit un résultat / se situe dans	*processus*	*résultat*	*localisation*

Contrairement aux relations conceptuelles de la règle [ϕ [x_1]] (v. *supra*, p. 151), les relations établies par la règle [[ϕ [x_1]] [y_1]] se composent toujours de deux étiquettes séparées par une barre oblique (/). La première étiquette fait référence à la relation établie entre la base et l'argument, c'est-à-dire celle qui se réalise à l'intérieur de la prédication nucléaire. La seconde étiquette, qui vient après la barre oblique, résume la relation qui existe entre le satellite et la prédication nucléaire.

La relation conceptuelle *agit sur / se situe dans* caractérise les UTC dont la base est un *processus* ou un *agent* qui *agit sur* un *argument-patient.* L'ensemble base-processus / base-agent se situe dans une *satellite-lieu*. Voici quelques exemples tirés de notre corpus :

- *processus* + *patient* + *localisation*
  ex. : *aclimatação de plantas no solo* (acclimatation des plantes au sol);
  « l'acclimatation *agit sur* les plantes »;
  « l'acclimatation des plantes *se situe* dans le sol ».

- *agent* + *patient* + *localisation*
  ex. : *inibidor de crescimento in vitro* (inhibiteur de croissance in vitro);
  « l'inhibiteur *agit sur* la croissance »;
  « l'inhibiteur de croissance *se situe* in vitro (en laboratoire) ».

La relation *est du type* / *est en état* désigne une *base-processus* qui a un *argument-propriété* et un satellite qui indique dans quel état ce processus est produit. Ainsi, l'UTC *recombinação homóloga in vivo* (recombinaison homologue in vivo) est une recombinaison qui a la propriété d'être homologue. La recombinaison homologue se produit donc dans un organisme vivant.

Finalement la relation *produit un résultat / se situe dans* englobe une *base-processus* qui produit un *argument-résultat*. Ce processus et son résultat sont situés dans un *satellite-lieu*. Prenons par exemple l'UTC *indução floral in vitro* (induction florale in vitro). Il s'agit d'un type d'induction qui produit des fleurs. L'induction florale se passe *in vitro* (en laboratoire).

4.2.3 [ϕ [y_1]]

La règle de formation [ϕ [y_1]] représente les UTC du type *base* + *satellite*. Même si peu d'UTC de notre corpus appartiennent à cette règle, celle-ci est représentative puisqu'elle démontre un mode de formation du domaine en question. En plus, elle semble se caractériser comme étant une structure dérivée de la règle [[ϕ [x_1]] [y_1]]. Ce fait renforce donc son importance comme mode de formation.

Comme la règle [[ϕ [x_1]] [y_1]], la règle [ϕ [y_1]]. se caractérise par une *prédication nucléaire* et une *prédication étendue*. Ce qui la distingue c'est l'absence d'argument dans la structure de la règle [ϕ [y_1]]. Dans ce cas, la *prédication nucléaire* est

formée seulement de la *base*. Cette base nominale, tant du type *active* que *statique*, joue les rôles syntaxique du *sujet* et pragmatique de *topique*. Quant aux fonctions sémantiques, les bases provenant de cette règle se caractérisent comme étant des *entités* ou des *processus*.

Au cours de cette section, nous allons voir que l'argument peut être interprété comme un élément qui peut avoir existé au moins dans la pensée du spécialiste. La présence du satellite à côté de la base caractérise la *prédication étendue*. Étant donné que l'absence de l'argument est la principale caractéristique de cette règle, c'est sur cet aspect que nous allons concentrer les interprétations de cette section.

4.2.3.1 Satellite

Les satellites sont le moyen par lequel il est possible d'ajouter de nouvelles caractéristiques à la prédication nucléaire prise comme un tout. Selon Dik, « The possibilities of adding a certain satellite to a nuclear predication are determined by features of the nuclear predication as a whole rather than by the predicate. [...] If these satellites were to be separately specified for each predicate, this would lead to an enormous redundancy in the lexicon [...] » (1989 : 193-194). C'est probablement la raison pour laquelle dans cette règle, tout comme dans la règle [[ϕ [x_1]] [y_1]], ils occupent une position périphérique, c'est-à-dire qu'ils viennent après la prédication nucléaire.

Les satellites de la règle [ϕ [y_1]] appartiennent tous à la catégorie grammaticale *locution adverbiale*. Nous en avons recensés trois : *in situ*, *in vitro* et *in vivo*. Considérées comme des structures canoniques venues directement du latin, ces locutions sont des véritables *locutions adverbiales du domaine*. Il s'agit de locutions qui « [...] n'acceptent qu'un seul type de fonctionnement, le fonctionnement exophrastique » (Guimier, 1996 : 144). Autrement dit, ce sont des satellites classés comme des *locutions adverbiales* qui se trouvent toujours dans la position de prédication étendue. C'est justement la raison pour laquelle nous les appelons *locutions adverbiales étendues*.

Ce n'est cependant pas toujours le cas. Dans la règle [ϕ [x_1]], nous avons vu des *locutions adverbiales* qui fonctionnaient dans la prédication nucléaire et qui étaient, par conséquent, des arguments et non des satellites. En outre, nous allons voir, dans d'autres règles, qu'il y a des adverbes du domaine qui utilisent le suffixe *-ment*, mais qui ne sont pas des satellites et qui n'occupent pas une position périphérique dans l'UTC.

Tous les satellites de la règle présentement analysée ont la fonction syntaxique *complément circonstanciel* et la fonction pragmatique *focus*.

Quant aux fonctions sémantiques, les satellites représentent essentiellement l'*état* et la *localisation.* Nous avons réuni quelques exemples de fonctions sémantiques dans le tableau suivant.

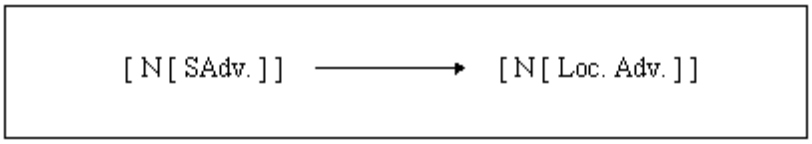
Tableau 4.20 — Fonctions sémantiques des satellites pour la règle [ϕ [y_1]]

Fonction sémantique des satellites	**UTC**
état	*antera* ***in vivo*** (anthère in vivo)
localisation	*hibridação* ***in situ*** (hybridation in situ) *mutagênese* ***in vitro*** (mutagenèse in vitro)

4.2.3.2 Analyse morphosyntaxique

La règle [ϕ [y_1]] présente une structure syntaxique du type [N [SAdv.]] qui prend la formule morphologique [N [Loc. Adv.]] comme illustré dans la figure 4.10.

Figure 4.10 — Formule syntaxique et représentation en formule morphologique pour la règle [ϕ [y_1]]



Voici les exemples trouvés pour ce type de structure :

- UTC[N *alongamento* Loc. Adv.[*"in vitro"*]] (élongation "in vitro");
- UTC[N *antera* Loc. Adv.[*in vivo*]] (anthère in vivo);
- UTC[N *cultivo* Loc. Adv.[*"in vitro"*]] (culture "in vitro");
- UTC[N *cultura* Loc. Adv.[*in vitro*]] (culture in vitro);
- UTC[N *embriogênese* Loc. Adv.[*in vitro*]] (embryogenèse in vitro);
- UTC[N *experimento* Loc. Adv.[*in vitro*]] (expérimentation in vitro);
- UTC[N *hibridação* Loc. Adv.[*in situ*]] (hybridation in situ);
- UTC[N *método* Loc. Adv.[*in vitro*]] (méthode in vitro);
- UTC[N *mutagênese* Loc. Adv.[*in vitro*]] (mutagenèse in vitro);
- UTC[N *mutagênese* Loc. Adv.[*in vivo*]] (mutagenèse in vivo);
- UTC[N *polinização* Loc. Adv.[*in vitro*]] (pollinisation in vitro);
- UTC[N *propagação* Loc. Adv.[*in vitro*]] (propagation in vitro);
- UTC[N *regeneração* Loc. Adv.[*in vitro*]] (régénération in vitro).

L'analyse syntaxique de ces exemples nous amène à conclure que les locutions adverbiales de cette règle prédiquent la base nominale. Étant donné que la base nominale représente la prédication nucléaire, nous pouvons supposer que ces locutions portent sur l'intégralité de la prédication nucléaire et la modélisent de l'extérieur, comportement que Guimier appelle extra-prédicatif (v. 1996, *supra*, p. 34).

4.2.3.3 Relation conceptuelle entre la base et le satellite

Les relations conceptuelles présentées dans le prochain tableau découlent des rapports établis entre les fonctions sémantiques de la base et du satellite de la règle [ϕ [y_1]].

Tableau 4.21 — Relations conceptuelles et fonctions sémantiques pour la règle [ϕ] [y_1]]

Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**	
	base	**satellite**
se situe dans	*processus* *entité*	*localisation* *localisation*
est en état	*processus* *entité*	*état* *état*

La relation conceptuelle *se situe dans* représente les UTC dont la base est un *processus* ou une *entité* qui *se situe dans* un *satellite-lieu*. La relation *est en état* désigne pour sa part une *base-processus* ou une *base-entité* que *est en état* désigné par un *satellite-état*. Examinons quelques exemples :

- *processus* + *localisation*
  ex. : *hibridação in situ* (hybridation in situ);
  « l'hybridation *se situe* in situ (dans son milieu naturel) ».

- *entité* + *localisation*
  ex. : *método in vitro* (méthode in vitro);
  « la méthode *se situe* in vitro (en laboratoire) ».

- *processus* + *état*

ex. : *mutagênese in vivo* (mutagenèse in vivo);
« la mutagenèse *est en état* de vie (en conditions naturelles) ».

- *entité + état*
  ex. : *antera in vivo* (anthère in vivo);
  « l'anthère *est en état* de vie ».

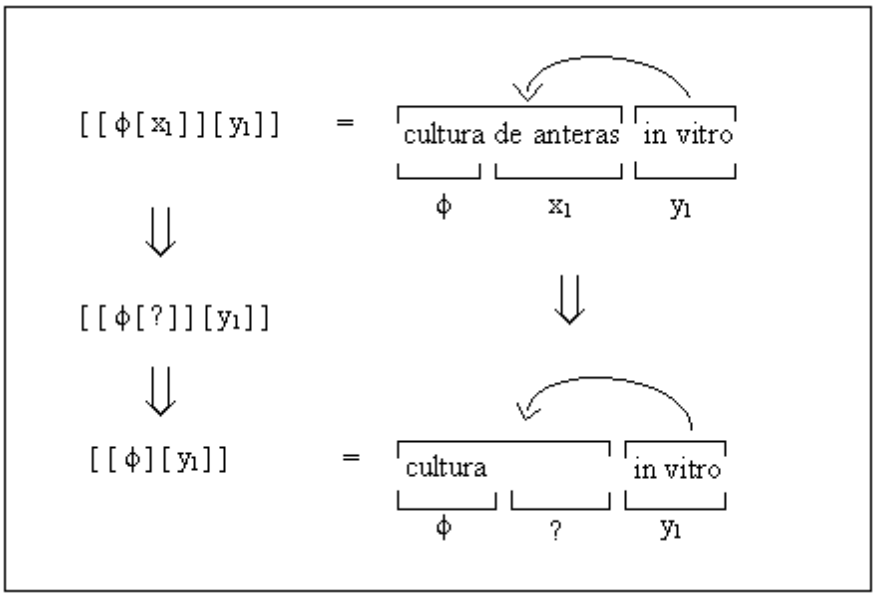
Nous remarquons que dans tous les cas où la base est un *processus* et le satellite est un *état* (*in vivo*), il semble y avoir un effacement d'un argument ou d'une partie du satellite. L'absence de ces composants rend la compréhension des rapports entre les fonctions difficile. Par exemple, l'UTC *mutagênese in vivo* (mutagenèse in vivo) exprime, comme nous l'avons vu, une mutagenèse qui *est en état* de vie. L'analyse de *in vivo* est facile à faire dans des UTC composées d'une *base-entité* comme *antera in vivo* (anthère in vivo) « anthère qui *est en état* de vie ». Toutefois, dans le cas de *mutagênese in vivo* (mutagenèse in vivo), les choses se compliquent un peu puisque ce n'est pas exactement la mutagenèse qui *est en état* de vie, mais un *patient* du processus de mutagenèse (sous-entendu dans l'UTC) qui est en état vivant. Par contre, si nous acceptons la définition de l'expression *in vivo* présentée dans *Le Nouveau Petit Robert*, soit « en conditions naturelles en opposition à in vitro (en laboratoire) » (v. 1995), ce satellite pourrait également avoir un rapport direct avec la *base-processus*. Par conséquent, *in vivo* (en conditions naturelles) peut s'appliquer à :

a) un *satellite-état* qui a un rapport avec une *base-processus* (*mutagênese in vivo*);
b) un *satellite-état* qui a un rapport avec une *base-entité* (*antera in vivo*);
c) un *satellite-état* qui a un rapport avec une *prédication nucléaire* prise comme un tout et composée d'une *base-processus* et un *argument-patient* sous-entendu (**mutagênese ? in vivo*).

La comparaison entre les UTC de la règle[ϕ] [y_1] et celles de la règle [[ϕ [x_1]] [y_1], qui comportent une *base-processus* et le *satellite-localisation in vitro*, nous laisse

supposer que la règle [ϕ] [y_1]] dérive de la règle [[ϕ [x_1]] [y_1]]. Tel est le cas de *cultura de anteras in vitro* (culture d'anthères in vitro) et *cultura in vitro* (culture in vitro), illustrées ci-dessous. Si l'on compare les deux règles, on s'aperçoit que, dans la règle [ϕ] [y_1], il y a eu effacement de l'argument *antera* (anthère) qui était présent dans la règle [[ϕ [x_1]] [y_1].

Figure 4.11 — Dérivation de la règle [ϕ] [y_1]] à partir de la règle [[ϕ [x_1]] [y_1]



Des analyses similaires pourraient être effectuées pour l'ensemble d'UTC citées ci-dessous :

- *propagação clonal in vitro* (propagation clonale in vitro) ⇒ *propagação in vitro* (propagation in vitro);
- *propagação vegetativa in vitro* (propagation végétative in vitro) ⇒ *propagação in vitro* (propagation in vitro);
- *método de propagação in vitro* (méthode de propagation in vitro) ⇒ *método in vitro* (méthode in vitro).

Il reste à voir si ce processus apparaît seulement dans des conditions discursives ou s'il s'agit des modes de formation lexicalisées. Nous aurions tendance à croire qu'il s'agit plutôt de structures figées puisque nous en avons trouvé quelques exemples dans des ouvrages de référence du domaine. Ainsi, « culture in vitro » a été repéré dans Termium et « hybridation in situ » dans LSB et Termium.

4.2.4 [ϕ [x_1 [$x_{1/1}$]]]

Les UTC de la règle [ϕ [x_1 [$x_{1/1}$]]] sont formées d'une *prédication nucléaire*, c'est-à-dire d'une *base* + un *argument*. Toutefois, contrairement à la règle [ϕ [x_1]], celle-ci se caractérise par une expansion de l'argument. Ce comportement a déjà été noté par d'autres auteurs, notamment par Kocourek qui mentionne que

> La formule syntaxique linéaire ne signale pas la structure hiérarchique des rapports internes cachés. Dans les syntagmes nominaux subordinatifs, le régissant est constitué d'un nom, qui est modifié par divers modificateurs, tels qu'un adjectif, un syntagme prépositionnel (avec un nom ou un infinitif), un syntagme nominal, ou une combinaison des précédents. Le modificateur peut être lui-même modifié, étendu (1991 : 141).

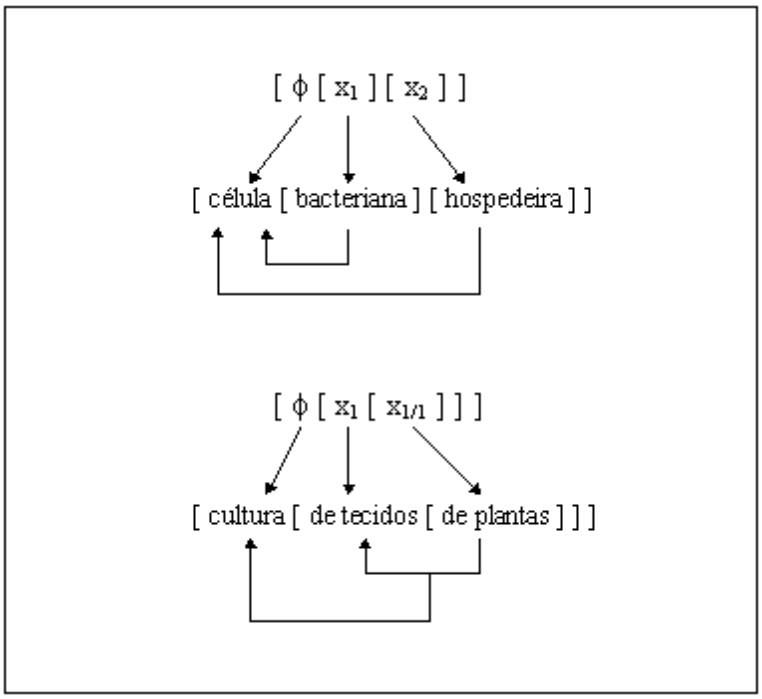
C'est justement sur l'argument modificateur d'un autre argument que nous allons nous pencher dans la section qui suit.

4.2.4.1 Argument et expansion

Tout d'abord, il faut préciser que l'expansion à droite dans le cas de la règle [ϕ [x_1 [$x_{1/1}$]]] est différente de celle des règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] que nous verrons plus loin dans ce chapitre. Dans la première règle, c'est l'argument qui est modifié, tandis que dans les deux autres c'est la base qui reçoit plusieurs arguments-modificateurs. Nous pouvons vérifier cette distinction dans les UTC *célula bacteriana hospedeira* (cellule

bactérienne hôte) et *cultura de tecidos de plantas* (culture de tissus de plantes) illustrées dans la figure qui suit. La première UTC est du type [ϕ [x_1] [x_2]], tandis que la seconde découle de la structure [ϕ [x_1 [$x_{1/1}$]]].

Figure 4.12 — Rapports selon le type de relation entretenue par l'argument



Dans le cas de *célula bacteriana hospedeira* (cellule bactérienne hôte), c'est la cellule qui se situe dans la bactérie et qui agit comme un hôte. Nous voyons bien que le rapport des arguments se fait toujours avec la base *célula* (cellule). Ce qui n'est pas le cas pour les arguments de *cultura de tecidos de plantas* (culture de tissus de plantes) où c'est l'ensemble *tecidos de plantas* (tissus de plantes) qui est le résultat du processus de *cultura* (culture). L'argument *plantas* (plantes) maintient un rapport avec *tecidos* (tissus) et non avec *cultura* (culture). En effet, ce sont les tissus de plantes qui sont produits par la culture et non les plantes en soi.

Le phénomène représenté par la règle [ϕ [x_1 [$x_{1/1}$]]] caractérise une espèce d'enchaînement d'UTC, autrement dit, il s'agit d'une association d'UTC pour former le

concept désiré par le spécialiste du domaine. Dans la structure de la règle, ce lien est illustré par la barre oblique (/). L'élément $x_{1/1}$ signifie la première expansion du premier argument. Ainsi, le chiffre qui est à droite de la barre indique l'expansion, tandis que celui qui est à gauche indique l'argument qui est modifié.

4.2.4.2 Argument-adverbe et expansion

Le cadre que nous venons de voir semble relativement simple à visualiser. Toutefois, des exemples plus complexes ont été recueillis dans notre corpus. Tel est le cas des UTC composées d'un adverbe suivi d'un adjectif. En effet, l'adverbe est une catégorie grammaticale qui, jusqu'à maintenant, n'a pas été étudiée en profondeur en terminologie. Dans un article récent, nous avons montré que même si l'adverbe est peu souvent pris en considération dans les listes de patrons morphosyntaxiques, il entre quand même dans les modes de formation des UTC en portugais, en français et en anglais (v. Miranda et Cochrane, 1997).

Dans le cas de la règle [ϕ [x_1 [$x_{1/1}$]]], les adverbes sont, au niveau syntaxique, des compléments circonstanciels qui, avec l'adjectif en fonction d'adjoint de nom, précisent le signifié de la base-sujet. Même si, comme l'affirme Guimier, le contenu de l'adverbe « dépend étroitement des autres éléments du prédicat auquel il appartient » (1996 : 2), nous croyons que son rapport direct se fait avec son expansion. En effet, c'est l'ensemble argument-adverbe et expansion-adjectif qui entretient un rapport avec la base et non chaque composant pris séparément comme c'est le cas pour la règle [ϕ [x_1] [x_2]].

Sur le plan fonctionnel, ces composants sont des *adverbes nucléaires* en raison de la place qu'ils occupent dans la structure de la prédication. Dans notre corpus, nous avons trouvé deux types d'adverbes dont le rôle est d'être un argument et qui font donc partie de la prédication nucléaire. Le premier type se réfère aux adverbes formés d'adjectifs du domaine plus du suffixe *-mente* (-ment*)*. Le second type comporte des adverbes de la langue générale

qui finissent également en *-mente* (-ment). Nous avons regroupé, dans le tableau qui suit, des exemples de ces deux ensembles.

Tableau 4.22 — Exemples d'UTC formées d'arguments adverbiaux

Adverbes du domaine	**Adverbes de la langue générale**
• *cloroplasto* ***fotossinteticamente*** *competente* (chloroplaste photosynthétiquement compétent) • *espécie sexualmente compatível* (espèce sexuellement compatible) • *espécies* ***sexualmente*** *incompatíveis* (espèces sexuellement incompatibles) • *indivíduos* ***genotipicamente*** *idênticos* (individus génotypiquement identique) • *planta* ***genotipicamente*** *uniforme* (plante génotypiquement uniforme) • *regenerante* ***geneticamente*** *engenheirado* (régénérant génétiquement modifié) • *regenerante* ***geneticamente*** *estável* (régénérant génétiquement stable)	• *DNA* ***altamente*** *repetitivo* (ADN hautement répétitif) • *espécie* ***altamente*** *endógama* (espèce hautement endogame) • *mRNA* ***altamente*** *purificado* (ARNm hautement purifié) • *tecido* ***altamente*** *diferenciado* (tissu hautement différencié)

Éléments invariables, ces adverbes suivis d'un adjectif ont pour fonction de modifier la base de l'UTC. L'adverbe ne qualifie pas, il spécifie la qualification fournie par l'adjectif. Ainsi, dans l'UTC *cloroplasto* ***fotossinteticamente*** *competente* (chloroplaste photosynthétiquement compétent), c'est l'expansion adjectivale *competente* (compétent) qui qualifie la base nominale *cloroplasto* (chloroplaste). L'argument adverbial *fotossinteticamente* (photosynthétiquement) spécifie le type de compétence que le chloroplaste possède.

Étant donné que peu d'auteurs se sont intéressés aux adverbes dans les UTC, il est difficile d'établir une classification sémantique de la spécification fournie par l'adverbe. Si on regarde les grammaires, on constate que la typologie proposée ne s'applique pas toujours aux UTC. L'adverbe de manière est l'un des types les plus complexes. Derrière l'étiquette

« manière » se cache des nuances importantes pour la compréhension des rapports sémantiques entre les composants des UTC. Dans son analyse des adverbes de manière, Guimier mentionne que « Sur le plan sémantique, la complémentation permise par l'adverbe intra-prédicatif appartient à des domaines variés, dont la manière ne constitue qu'un cas particulier, aux contours d'ailleurs difficilement définissables » (1996 : 59).

Dans les exemples présentés dans le tableau précédent, les adverbes marquent trois types de fonctions sémantiques. Les adverbes des UTC de la colonne de gauche sont interprétés sémantiquement comme des *instruments* ou des *composants* tandis que ceux appartenant aux UTC placées dans la colonne de droite sont des adverbes d'*intensité*. Nous allons maintenant expliquer ces deux types d'étiquettes.

Les adverbes dits instrumentaux sont ceux qui lient une *entité* ou un *agent* à une *propriété*, un *état* ou un *effet*. Prenons encore une fois l'exemple de *cloroplasto* ***fotossinteticament****e competente* (chloroplaste photosynthétiquement compétent). La base *cloroplasto* (chloroplaste) est l'*agent* qui produit l'effet de la photosynthèse. Alors l'argument ***fotossinteticament****e* (photosynthétiquement) est l'*instrument* de l'*effet* de photosynthèse produit par l'*agent cloroplasto* (chloroplaste).

Les adverbes qui expriment un *composant* reflètent un rapport entre une *entité* et une *propriété* ou un *état*. Ainsi, dans l'UTC *planta* ***genotipicamente*** *uniforme* (plante génotypiquement uniforme), l'adverbe *genotipicamente* (génotypiquement) précise que la plante est composée de génotypes (d'un ensemble de gènes) uniformes. Si on considère que génotype est défini comme « un patrimoine génétique », on pourrait dire que cette UTC fait référence à une plante composée d'un ensemble uniforme de gènes. Par conséquent, l'argument *genotipicamente* indique le *composant* de l'uniformité qui est la propriété de la plante.

Fait intéressant à noter, cette formation adverbiale est attestée dans d'autres termes du domaine relevés dans les principales banques de terminologie au Canada. Tels sont le

cas d'« organisme génétiquement modifié » (Termium) et de « rayonnement photosynthétiquement actif » (BTQ). Ces exemples semblent définir une tendance dans la formation des UTC de la biotechnologie des végétaux et s'observent dans d'autres langues que le portugais.

Dans les UTC analysées, les adverbes du type *intensité* peuvent prendre un sens qualificatif ou quantitatif. Ainsi, dans l'UTC *DNA* ***altamente*** *repetitivo* (ADN hautement répétitif), l'adverbe a pour fonction d'intensifier la quantité de « séquences d'ADN qui sont présentes en copies multiples dans chaque jeu de chromosomes » (v. Quirion et Bourbeau, s.d.). Par contre, dans l'UTC *espécie* ***altamente*** *endógama* (espèce hautement endogame), l'adverbe intensifie la qualité de l'espèce qui pratique un « mode de reproduction sexuée par fécondation entre deux gamètes provenant d'un même individu ou entre individus apparentés » (v. Le Nouveau Petit Robert, 1995). Il en est de même pour l'UTC *tecido* ***altamente*** *diferenciado* (tissu hautement différencié) où l'adverbe vient intensifier qualitativement la propriété du tissu d'être différenciée. D'ailleurs, cette propriété existe déjà dans le concept de tissu comme on le voit dans la définition de tissu donnée par le *Novo dicionário da língua portuguesa* : « ensemble de cellules d'origine commune également différenciées pour réaliser certaines fonctions dans un organisme vivant » (v. 1986) [54].

Feuillet, qui a tenté d'établir une typologie des adverbes, propose une classification basée sur des critères fonctionnels. Les adverbes intensifs feront partie de la classe des adverbes de fonctions modificatrices « qui ne portent que sur un groupe ou un noyau (et qui n'exprime pas de jugement sur le *dictum*) » (1991 : 49).

Certains adverbes sont dérivés d'adjectifs. On remarque cette caractéristique dans les exemples qui comportent un adverbe du domaine. Ainsi, pour l'UTC *cloroplasto* ***fotossinteticament****e competente* (chloroplaste photosynthétiquement compétent) l'adverbe est formé à partir de l'adjectif *fotossintético* (photosynthétique). Toutefois, dans le cas des

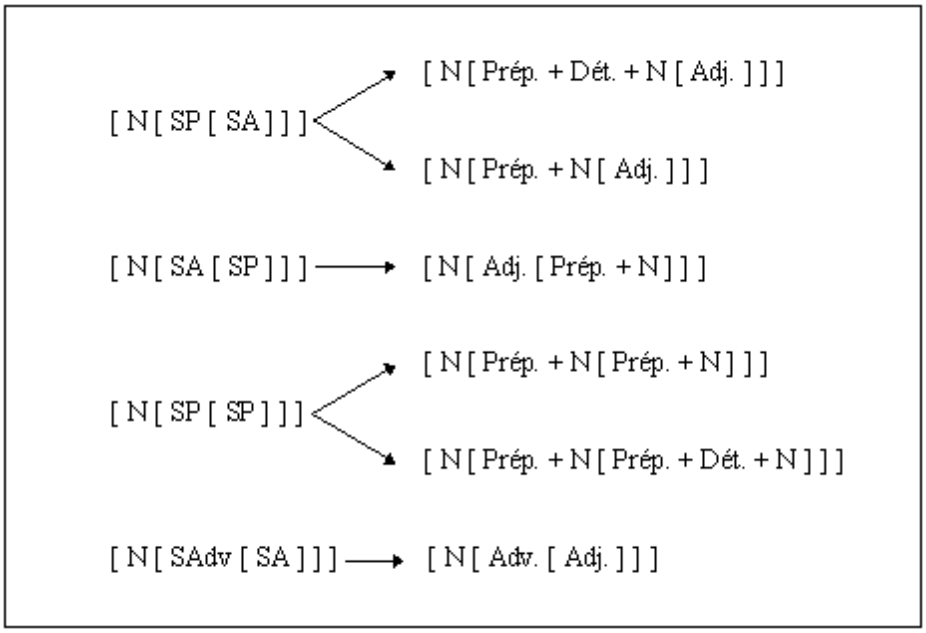
intensifs, cette transformation s'avère impossible. Comme l'explique Feuillet, « [...] les adverbes exerçant des fonctions modificatrices n'ont jamais de correspondants exacts parmi les adjectifs, qu'il s'agisse des gradatifs, des intensifs, des restrictifs ou des spécificatifs [...] » (id. : 56).

Nous avons également observé que le sens de certaines UTC ne se distingue que par l'adverbe qu'elles comportent. Tel est le cas des UTC *tecido* ***altamente*** *diferenciado* (tissu hautement différencié) et *tecido diferenciado* (tissu différencié).

4.2.4.3 Analyse morphosyntaxique

La présence des expansions et des adverbes dans les UTC formées selon la règle [ϕ [x_1 [$x_{1/1}$]]], donne lieu à de nouvelles formules syntaxiques et morphologiques que nous avons regroupées dans la figure ci-dessous.

Figure 4.13 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1 [$x_{1/1}$]]]



[54] Traduction personnelle du texte original en portugais.

Afin de bien comprendre ces nouvelles formules morphologiques, nous avons réunis quelques exemples d'UTC dans le tableau qui suit.

Tableau 4.23 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1 [$x_{1/1}$]]]

Formule morphologique	**UTC**
[N [Prép. + Dét. + N [Adj.]]]	*mecanismo da*[55] *ação hidrolítica* (mécanisme de l'action relatif à l'hydrolyse)
[N [Prép. + N [Adj.]]]	*bactéria de vida livre* (bactérie à vie autonome)
[N [Adj. [Prép. + N]]]	*bactéria fixadora de nitrogênio* (bactérie fixatrice d'azote)
[N [Prép. + N [Prép. + N]]]	*vetor de clonagem de genes* (vecteur de clonage de gènes)
[N [Prép. + N [Prép. + Dét. + N]]]	*sistema de preparo do*[56] *solo* (système de préparation du sol)
[N [Adv. [Adj.]]]	*cloroplasto fotossinteticamente competente* (chloroplaste photosynthétiquement compétent)

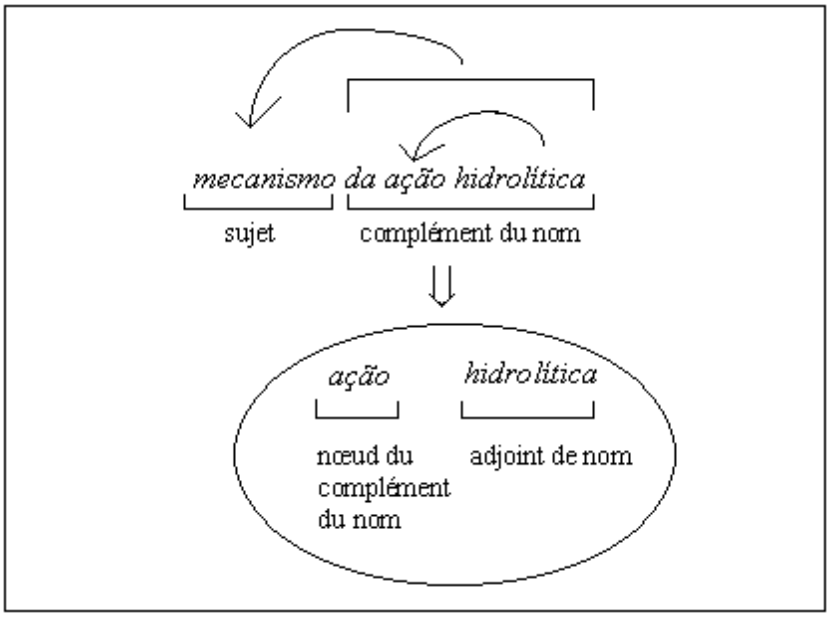
Il est intéressant d'observer que deux des formules morphologiques présentées ci-dessus comportent des déterminants. D'ailleurs, nous avons constaté que pour l'exemple *sistema de preparo* ***do*** *solo* (système de préparation du sol), il existe en français un équivalent similaire qui contient lui aussi l'article contracté, soit « préparation **du** sol » (v. Termium).

Le comportement prédicatif de chaque UTC est perçu non seulement dans la règle de formation, mais également dans les formules syntaxiques et morphologiques. Le comportement syntaxique reflète les fonctions sémantiques et pragmatiques exprimées par chaque prédicat. Ainsi, dans l'UTC *mecanismo da ação hidrolítica* (mécanisme de l'action relative à l'hydrolyse), nous remarquons que *mecanismo* (mécanisme) a la fonction de sujet

[55] *Da* est la contraction de la préposition *de* + de l'article féminin *a*.
[56] *Do* est la contraction de la préposition *de* + de l'article masculin *o*.

et que l'ensemble *ação hidrolítica* (action relative à l'hydrolyse) est le complément du nom. L'expansion *hidrolítica* (relative à l'hydrolyse) a la fonction d'adjoint de nom de *ação* (action). Pour sa part, *ação* (action) est le nœud du complément du nom. L'analyse de cette UTC est illustrée dans la figure 4.14.

Figure 4.14 — Fonctions syntaxiques pour la règle [ϕ [x_1 [$x_{1/1}$]]]



4.2.4.4 Rôle de l'argument-expansion dans les relations conceptuelles

Les étiquettes qui représentent les relations conceptuelles de la règle [ϕ [x_1 [$x_{1/1}$]]] symbolisent les rapports entre la base (ϕ) et l'ensemble argument + expansion (x_1 [$x_{1/1}$]). Dans la mesure où les expansions ($x_{1/1}$) n'ont pas un rapport direct avec la base, mais plutôt indirect, via l'argument, c'est l'ensemble argument + expansion qui est pris comme élément d'analyse. Cependant, les expansions sont analysées selon leurs fonctions sémantiques à l'intérieur de l'ensemble argument + expansion. C'est pourquoi nous regroupons dans le tableau suivant les relations conceptuelles suivies des fonctions sémantiques des trois composants de cette règle.

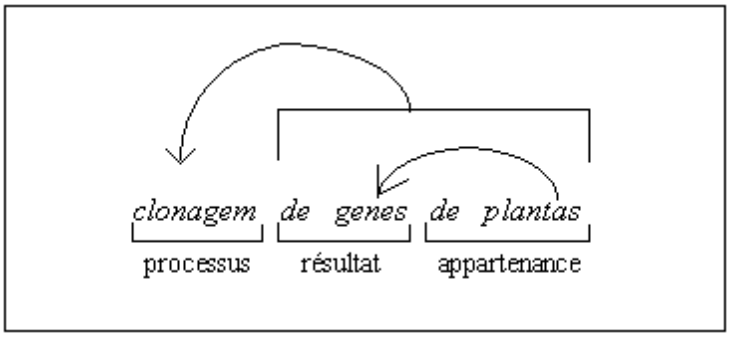
Tableau 4.24 — Relations conceptuelles et fonctions sémantiques pour la règle[ϕ [x_1 [$x_{1/1}$]]]

Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**		
	base	**argument**	**expansion de l'argument**
agit de manière à	*agent*	*propriété*	*propriété*
agit sur	*agent*	*patient*	*appartenance*
appartient	*entité* *entité* *entité*	*appartenance* *appartenance* *appartenance*	*état* *mode* *instrument*
composé de	*entité* *entité*	*composant* *composant*	*état* *propriété*
conduit par	*processus*	*processus*	*appartenance*
contient	*entité* *entité* *entité* *entité* *groupement*	*contenance* *contenance* *contenance* *contenance* *contenance*	*appartenance* *processus* *propriété* *contenance* *appartenance*
est du type	*entité* *entité* *entité*	*instrument* *intensité* *processus*	*propriété* *propriété* *état*
produit un effet	*agent* *agent* *agent* *processus*	*effet* *effet* *instrument* *effet*	*instrument* *patient* *effet* *mode*
produit un résultat	*processus* *processus*	*résultat* *résultat*	*appartenance* *localisation*
se situe dans	*processus*	*localisation*	*processus*
subit l'action	*processus*	*action*	*mode*

En consultant ce tableau, on remarque que, dans la plupart des relations conceptuelles, ce sont les fonctions sémantiques de la base et de l'argument qui expriment le rapport. Ainsi, pour une relation *produit un résultat* il y a une base *processus* et un argument *résultat.* Dans cette relation l'expansion peut être une *appartenance* ou une *localisation*, c'est-à-dire que c'est le *résultat* (et non le *processus*) qui peut appartenir ou

être situé. Revoyons cette analyse à l'aide d'un exemple concret. Dans l'UTC *clonagem de genes de plantas* (clonage de gènes de plantes), la base *clonagem* (clonage) est un *processus* qui produit comme *résultat* des *genes* (gènes). Ces gènes *appartiennent* à la plante. Nous reproduisons ces relations ainsi que les fonctions sémantiques dans la figure 4.15.

Figure 4.15 — Relation conceptuelle produit un résultat pour la règle[ϕ [x_1 [$x_{1/1}$]]]



Dans cette UTC, l'argument *genes* (gènes) est en même temps le *résultat* du processus de clonage et le *patient* puisque dans un processus de clonage, le gène est inséré dans une cellule hôte qui, au moyen d'un vecteur, va former une lignée de cellules qui possèdent ce gène (clone). Si on interprète cet argument comme un *patient*, la relation conceptuelle sera plutôt *agit sur*. Cet exemple nous signale à quel point il est difficile d'établir une classification fermée pour les relations conceptuelles et pour les fonctions sémantiques. Notre expérience nous a montré que plus une UTC a de composants, plus son interprétation sémantique est complexe.

4.2.5 [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]

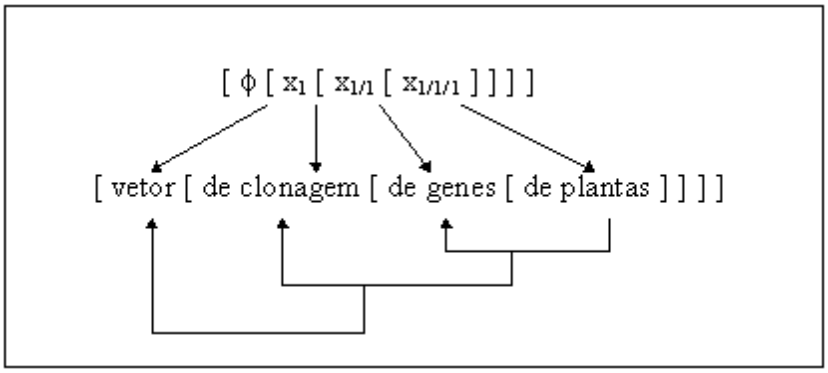
La règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] regroupe les UTC composées d'une *prédication nucléaire*. Comme la règle [ϕ [x_1 [$x_{1/1}$]]], elle possède une prédication nucléaire difficile à analyser. Cette difficulté découle des expansions successives à droite de l'argument. En fait, on pourrait dire qu'il s'agit d'un mode de formation qui prend comme base la règle [ϕ [x_1 [$x_{1/1}$]]]. Si la règle [ϕ [x_1 [$x_{1/1}$]]] représente les UTC dont l'argument est modifié,

celle-ci indique non seulement une expansion de l'argument mais aussi une modification de l'expansion de cet argument. C'est d'ailleurs sur ce point que nous allons concentrer les analyses dans la section qui suit.

4.2.5.1 Niveaux de rapports entre argument et expansions

Tout comme c'était le cas dans la règle [ϕ [x_1 [$x_{1/1}$]]], les expansions à droite de la base dans la présente règle se distinguent de celles des règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]]. Étant donné la complexité des rapports qui en découlent, nous avons une fois de plus eu recours aux crochets ([ ]) et à la barre oblique (/) pour systématiser l'analyse. Le mode de formation dont il est ici question caractérise le phénomène d'emboîtement des arguments associés les uns aux autres. Dans la figure qui suit, nous reproduisons l'UTC *vetor de clonagem de genes de plantas* (vecteur de clonage de gènes de plantes) pour illustrer le fait que nous venons d'expliquer.

Figure 4.16 — Rapports entre les composants de la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]



Dans cette figure, les flèches au-dessous de l'UTC ont la même fonction que les crochets et les barres obliques dans la représentation de la règle de formation. Les crochets encadrent l'ensemble des composants qui ont des rapports entre eux, tandis que les barres obliques indiquent quels sont les composants qui établissent un rapport donné. La lecture de

l'argument et des expansions se fait toujours de la droite vers la gauche. Les quatre composants de la règle analysée sont :

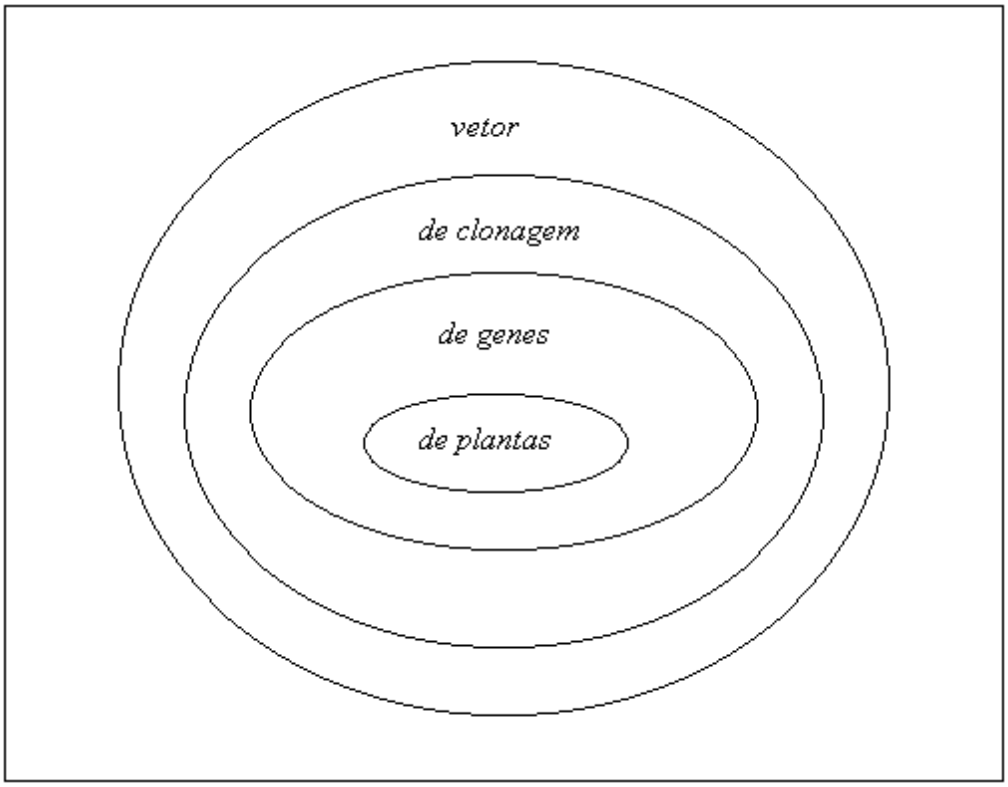
ϕ = base
x_1 = premier argument
$x_{1/1}$ = première expansion du premier argument
$x_{1/1/1}$ = première expansion de la première expansion du premier argument

À titre d'exemple, reprenons l'UTC *vetor de clonagem de genes de plantas* (vecteur de clonage de gènes de plantes) et examinons ses composants.

ϕ = base = *vetor*
x_1 = premier argument = *de clonagem*
$x_{1/1}$ = première expansion (*de genes*) du premier argument (*de clonagem) = de genes*
$x_{1/1/1}$ = première expansion (*de plantas*) de la première expansion (*de genes)* du premier argument (*de clonagem) = de plantas*

Les rapports établis entre l'argument et les expansions interprétés dans cet exemple de la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] sont de différents niveaux de dépendance. Ainsi, *de plantas* (de plantes) a un rapport direct avec *de genes* (de gènes), l'ensemble *de genes de plantas* (de gènes de plantes) entretient un rapport avec *de clonagem* (de clonage) et, finalement, l'ensemble *de clonagem de genes de plantas* (de clonage de gènes de plantes) établit la relation avec la base *vetor* (vecteur). Ces niveaux relationnels fonctionnent d'une manière hiérarchique qui met en rapport le modificateur et le modifié et sont représentés dans la figure suivante.

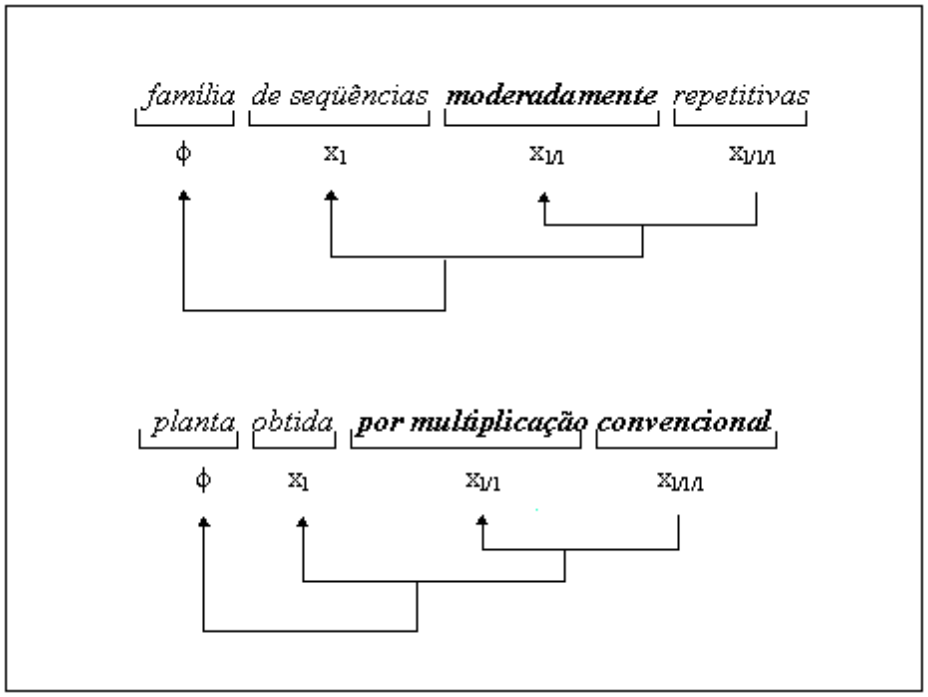
Figure 4.17 — Représentation des rapports entre les composants de la règle[ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]



4.2.5.2 Rôle syntaxique du complément circonstanciel

La fonction syntaxique *complément circonstanciel* apparaît de deux façons dans la structure de la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]], soit sous la forme d'adverbes, soit de syntagmes prépositionnels. Ce composant occupe toujours la troisième position ($x_{1/1}$) dans la structure et, avec l'adjectif, il modifie le premier argument. Analysons plus en détail, à l'aide de deux UTC présentées dans la figure ci-dessous, le comportement syntaxique du complément circonstanciel. L'UTC *família de seqüências* ***moderadamente*** *repetitivas* (famille de séquences modérément répétées) comporte un adverbe comme complément circonstanciel tandis que dans l'UTC *planta obtida* ***por multiplicação convencional*** (plante obtenue par multiplication conventionnelle), c'est un syntagme prépositionnel qui a cette fonction syntaxique.

Figure 4.18 — Représentation des compléments circonstanciels



Dans le premier exemple, l'adverbe *moderadamente* (modérément) signale une modification modale de la propriété des séquences d'être répétées. Dans l'axe paradigmatique, cette UTC a un rapport avec l'UTC *família de seqüências* ***altamente repetitivas*** (famille de séquences hautement répétées). En réalité, l'emploi de ces adverbes vient renforcer ou affaiblir une caractéristique déjà existante dans la notion de séquence répétée, c'est-à-dire une séquence d'organisation de l'ADN chromosomique qui se répète un grand nombre de fois (v. Quirion et Bourbeau, s.d.). Ce complément circonstanciel a une valeur adverbiale qui dénote, comme son nom l'indique, une circonstance de l'adjectif avec qui il a un rapport. Dans la structure de la prédication, ces adverbes sont interprétés comme des *adverbes nucléaires* puisqu'ils agissent à l'intérieur de la prédication nucléaire.

Dans le second exemple, la situation est semblable sauf que le complément circonstanciel est composé d'un syntagme prépositionnel (*por multiplicação*) et non d'un adverbe. Dans l'UTC *planta obtida* ***por multiplicação convencional*** (plante obtenue par

multiplication conventionnelle) l'adjectif *convencional* (conventionnelle), en fonction syntaxique d'adjoint de nom, spécifie le type de *multiplicação* (multiplication).

Pour ce qui est de la structure prédicative de l'UTC *planta obtida* ***por multiplicação convencional*** (plante obtenue par multiplication conventionnelle), le syntagme prépositionnel est une expansion de l'argument contenu dans la prédication nucléaire. Cette position à l'intérieur de la prédication nucléaire ainsi que sa fonction de locution adverbiale le classent, au plan fonctionnel, comme une *locution adverbiale nucléaire.*

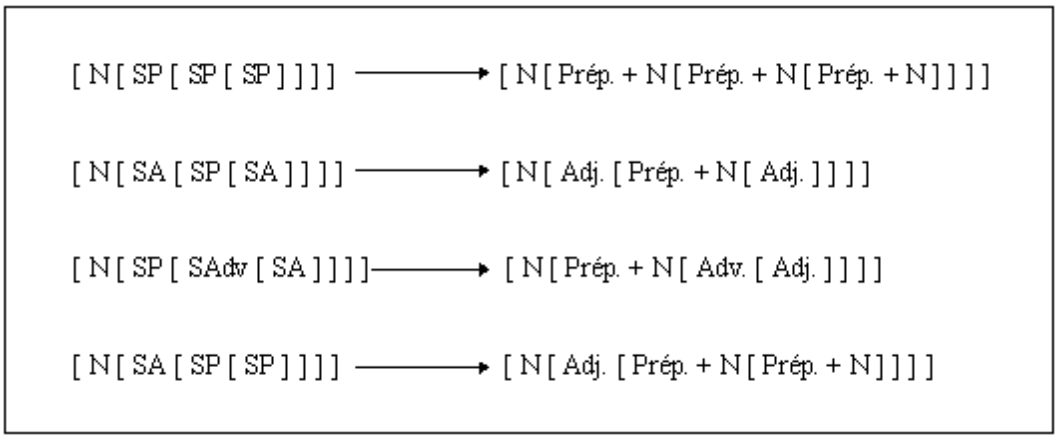
À première vue, cette UTC semble peu lexicalisée en raison de sa longueur et de sa structure. Toutefois, nous l'avons retenue pour deux raisons. Premièrement, dans le contexte, elle apparaît toujours suivie du terme *testemunha* (témoin) entre parenthèses ou vice versa, ce qui peut caractériser un degré de lexicalisation. Deuxièmement, la structure de cette UTC est reprise plusieurs fois dans les textes analysés et seuls la base et le premier argument sont alors substitués par des arguments équivalents. Voici quelques exemples retenus.

- *bulbilho obtido por multiplicação convencional* (bulbille obtenue par multiplication conventionnelle);
- *bulbo proveniente de multiplicação convencional* (bulbe provenant de la multiplication conventionnelle);
- *planta obtida por multiplicação convencional* (plante obtenue par multiplication conventionnelle);
- *planta originada de multiplicação convencional* (plante originaire de la multiplication conventionnelle);
- *planta oriunda de multiplicação convencional* (plante originaire de la multiplication conventionnelle);
- *planta proveniente de multiplicação convencional* (plante provenant de la multiplication conventionnelle).

4.2.5.3 Analyse morphosyntaxique

Pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]], nous avons relevé quatre types de patrons syntaxiques auxquels correspondent quatre patrons morphologiques. Tous ces patrons sont réunis dans la figure 4.19.

Figure 4.19 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]



En consultant cette figure, on remarque la présence de l'adverbe dans la formule syntaxique [N [SP [SAdv [SA]]]] et dans la formule morphologique correspondante [N [Prép. + N [Adv. [Adj.]]]]. Il faut également souligner la présence des locutions adverbiales. Celles-ci ont reçu l'étiquette syntagme prépositionnel (SP) puisque, dans notre étude, seules les expressions latines sont classées comme des locutions adverbiales. Ainsi, dans notre corpus, un SP englobe autant des séquences *De + Nom* que des locutions adverbiales. La consultation des exemples cités ci-dessous fait ressortir que l'UTC *bulbilho obtido* ***por multiplicação convencional*** (bulbille obtenue par multiplication conventionnelle) comporte un syntagme prépositionnel qui a une fonction modale de locution adverbiale, ce qui n'est pas le cas des autres exemples qui sont plutôt des syntagmes prépositionnels *De + Nom*.

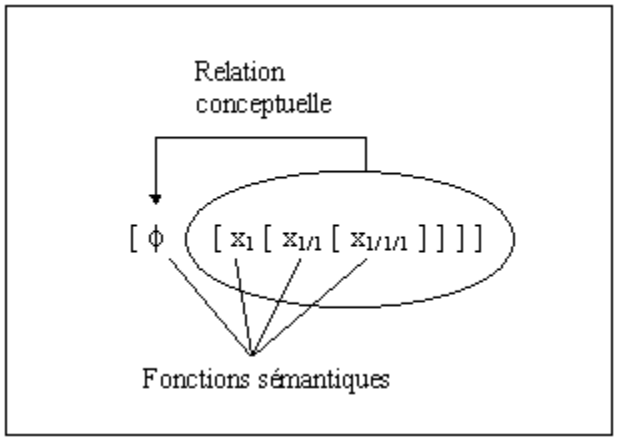
Tableau 4.25 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]

Formule morphologique	**UTC**
[N [Prép. + N [Prép. + N [Prép. + N]]]]	*vetor de clonagem de genes de plantas* (vecteur de clonage de gènes de plantes)
[N [Adj. [Prép. + N [Adj.]]]]	*bulbilho obtido por multiplicação convencional* (bulbille obtenue par multiplication conventionnelle)
[N [Prép. + N [Adv. [Adj.]]]]	*família de seqüências altamente repetitivas* (famille de séquences hautement répétées)
[N [Adj. [Prép. + N [Prép. + N]]]]	*método indexador de tipos de hospedeiros* (méthode d'indexage de types d'hôtes)

4.2.5.4 Rôle de l'argument-expansion dans les relations conceptuelles

Comme dans la règle [ϕ [x_1 [$x_{1/1}$]]], l'analyse des relations conceptuelles se fait entre la base et l'ensemble formé de l'argument et de ses expansions. De cette manière, même si nous avons plusieurs composants interreliés, nous attribuons une seule étiquette qui exprime la relation entre la base et l'ensemble des autres composants de l'UTC. Les relations qui s'établissent entre l'argument et les expansions sont indirectement perçues par leurs fonctions sémantiques. En effet, la relation conceptuelle est beaucoup plus proche de la définition de l'UTC, tandis que les fonctions sémantiques sont des éléments d'analyse nécessaires pour arriver à établir la relation conceptuelle de l'UTC prise comme un tout. Les principaux points de cette analyse sont illustrés dans la figure 4.20.

Figure 4.20 — Relation conceptuelle et fonctions sémantiques pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]



Nous avons recensé trois types de relations conceptuelles exprimées par les UTC de la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] Ces relations ainsi que les fonctions sémantiques correspondantes sont décrites dans le tableau suivant.

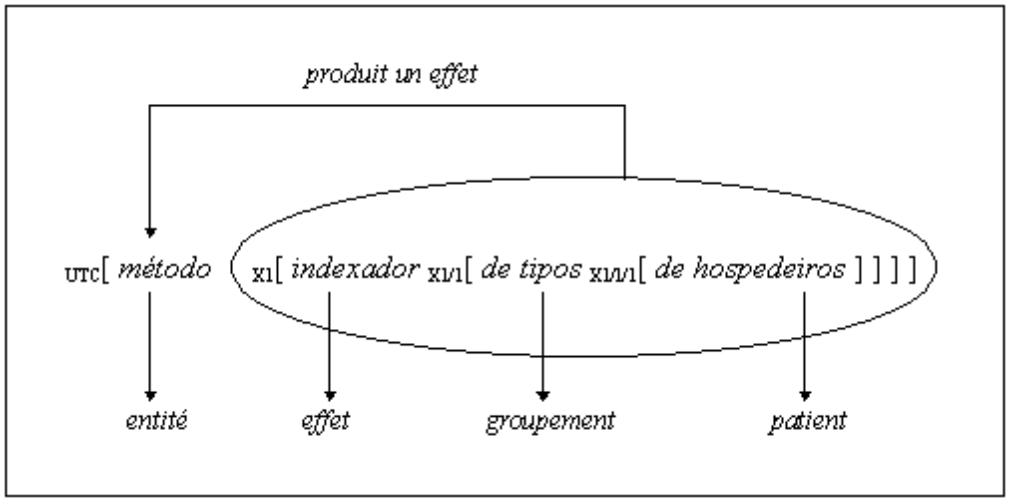
Tableau 4.26 — Relations conceptuelles et fonctions sémantiques pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]

Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**			
	base	**argument**	**expansion de l'argument**	**expansion de l'expansion de l'argument**
contient	*groupement*	*contenance*	*intensité*	*propriété*
produit un effet	*entité* *agent*	*effet* *effet*	*groupement* *patient*	*patient* *appartenance*
résulte de	*entité* *entité*	*source* *source*	*mode* *processus*	*mode* *mode*

L'UTC *família de seqüências altamente repetitivas* (famille de séquences hautement répétées) est un exemple de la relation *contient*. Cette UTC est formée d'une *base-groupement* (*família)* qui contient un *argument-contenance* (*de seqüências*). Cet *argument-contenance* a la propriété de se répéter très fréquemment.

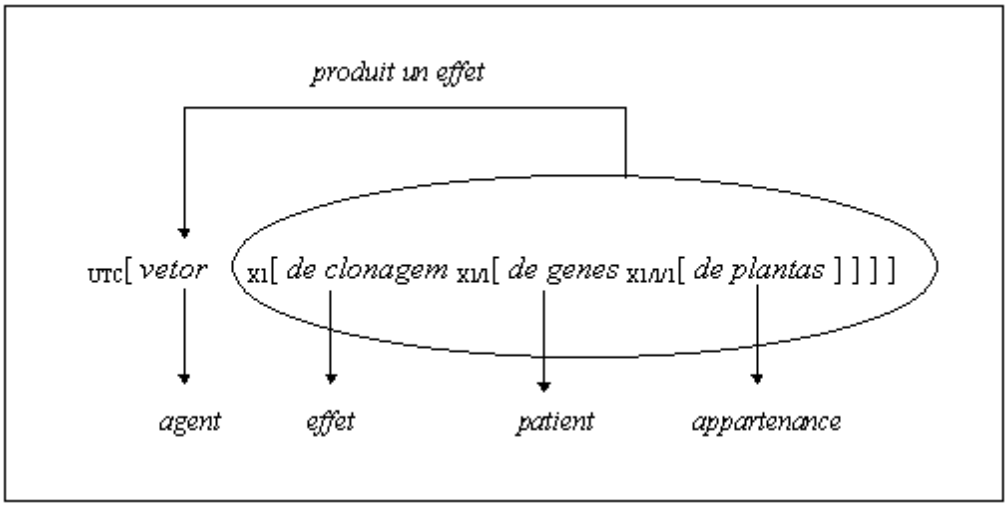
La relation *produit un effet* met en jeu deux séries de fonctions sémantiques pour les différents composants des UTC. L'UTC *método indexador de tipos de hospedeiros* (méthode d'indexage de types d'hôtes) illustre le premier groupe de fonctions sémantiques. Ainsi, *método* (méthode) est une *entité* qui *produit un effet* d'indexation. L'hôte est le *patient* du processus d'indexation et il est du type *groupement*. La figure suivante présente cette explication de façon schématique.

Figure 4.21 — Première relation conceptuelle produit un effet pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]



L'UTC *vetor de clonagem de genes de plantas* (vecteur de clonage de gènes de plantes) est, quant à elle, représentative du second groupe de fonctions sémantiques de la relation *produit un effet*. Dans cette UTC, le vecteur est un *agent* qui *produit un effet* de clonage. Les gènes sont des *patients* du *processus* de clonage et ils *appartiennent* aux plantes.

Figure 4.22 — Seconde relation conceptuelle produit un effet pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]

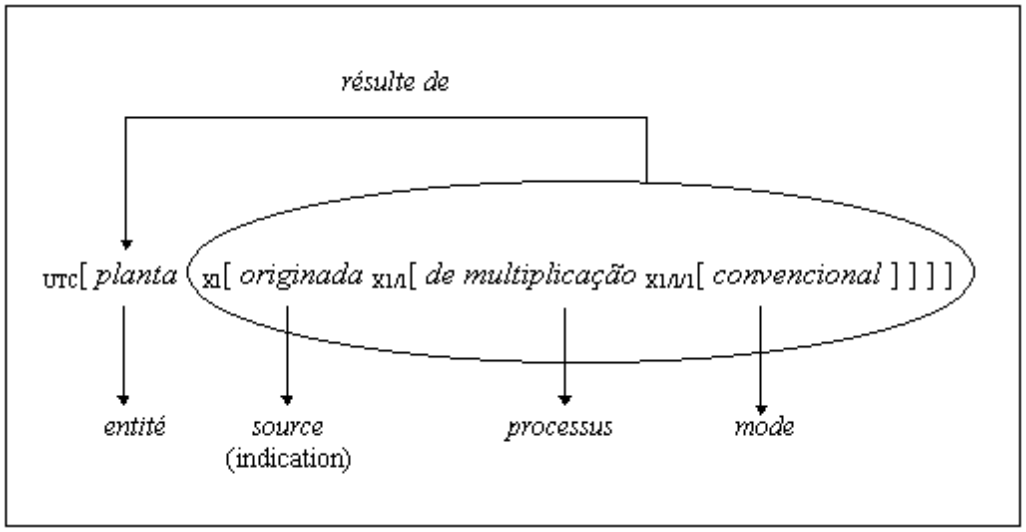


L'expansion *de genes* (de gènes) pourrait également être interprétée comme un *résultat* du *processus* de clonage produit par l'*agent* vecteur. C'est d'ailleurs de cette façon que nous avons analysé l'UTC *clonagem de genes de plantas* (clonage de gènes de plantes) de la règle [ϕ [x_1 [$x_{1/1}$]]] puisque, lorsqu'on regarde la définition de *clonagem*, on s'aperçoit que *de genes* (de gènes) est autant *patient* que *résultat*. Dans le *Glossário de Biotecnologia* clonage de gènes est défini comme un « processus par lequel un gène est inséré dans un vecteur et introduit dans une cellule-hôte qui, en se reproduisant, donnera naissance à un clone ou à une lignée de cellules qui abritent le gène étranger »[57].

La relation *résulte de* est un peu plus complexe à analyser. Les UTC qui ont reçu cette étiquette ont toujours une base *entité* suivie d'un argument qui indique la *source* exprimée par son expansion. La *source* est un type de *processus* et ce dernier est toujours spécifié par la dernière expansion qui a la fonction de *mode*. La figure qui suit montre l'analyse de l'UTC *planta originada de multiplicação convencional* (plante originaire de la multiplication conventionnelle).

[57] Traduction personnelle du texte original en portugais.

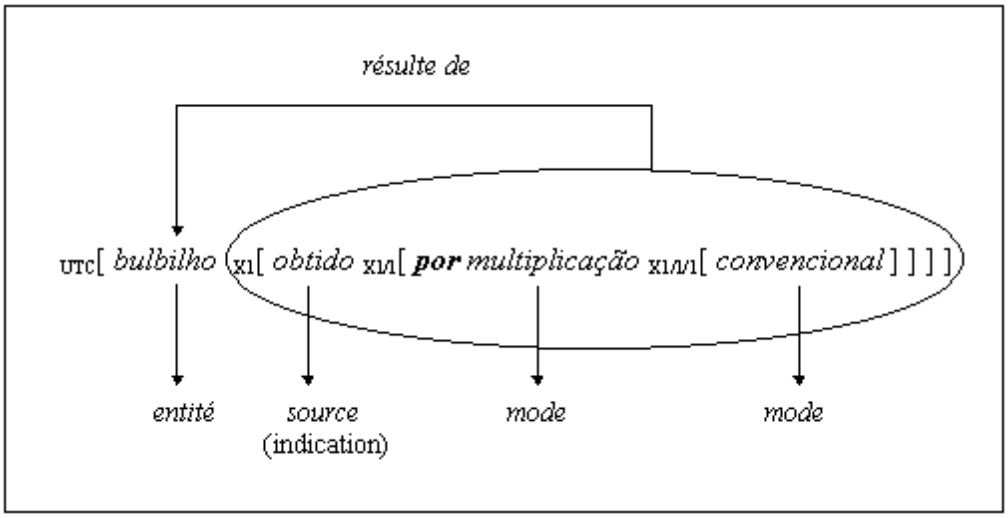
Figure 4.23 — Relation conceptuelle résulte de pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]



Comme on peut le constater, l'argument *originada* (originaire) n'est pas exactement la source en soi, mais plutôt une indication de l'existence d'une *source*. La source sera déduite par le biais de l'expansion de cet argument. Dans cette UTC, la fonction sémantique source est exprimée par le *processus de multiplicação* (de multiplication).

Le *processus* peut également être interprété comme un *mode*. Dans ce cas, il y aura un changement de la préposition. Tel est le cas de l'UTC *bulbilho obtido* ***por*** *multiplicação convencional* (bulbille obtenue par multiplication conventionnelle) qui est schématisée dans la figure ci-dessous.

Figure 4.24 — Changement de fonction sémantique dans la relation conceptuelle résulte de pour la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]



4.2.6 [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]]

Les UTC des règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] sont formées d'une *prédication nucléaire*. Cette prédication nucléaire est constituée d'une base et de plusieurs arguments. Pour ces règles, il s'agit vraiment d'arguments et non d'expansions d'arguments comme c'était le cas pour les règles [ϕ [x_1 [$x_{1/1}$]]], [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] et d'autres modes de formation que nous avons appelé des cas particuliers et qui seront présentés plus loin dans ce chapitre.

L'objet de cette section sera d'approfondir l'analyse des caractéristiques qui distinguent les règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] des autres modes de formation recueillis dans notre corpus.

4.2.6.1 Base et arguments

La principale caractéristique des règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] est que chaque argument a un rapport direct avec la base de l'UTC. C'est précisément ce qui les distinguent des autres règles qui comportent plus d'un composant à la droite. Afin de bien comprendre ces rapports, nous présentons ci-dessous une comparaison, sous forme d'arbres syntagmatiques, de ces deux règles avec la règle [ϕ [x_1 [$x_{1/1}$]]]. Cette comparaison sert à bien faire ressortir les rapports entre les composants des UTC *bactéria de vida livre* (bactérie à vie autonome), *árvore elite adulta* (arbre élite adulte) et *molécula de DNA circular extracromossômica* (molécule d'ADN circulaire extrachromosomique) et à montrer que celles-ci ne s'analysent pas de la même façon, quoiqu'on puisse en penser.

Figure 4.25 — Représentation de la règle [ϕ [x_1 [$x_{1/1}$]]]

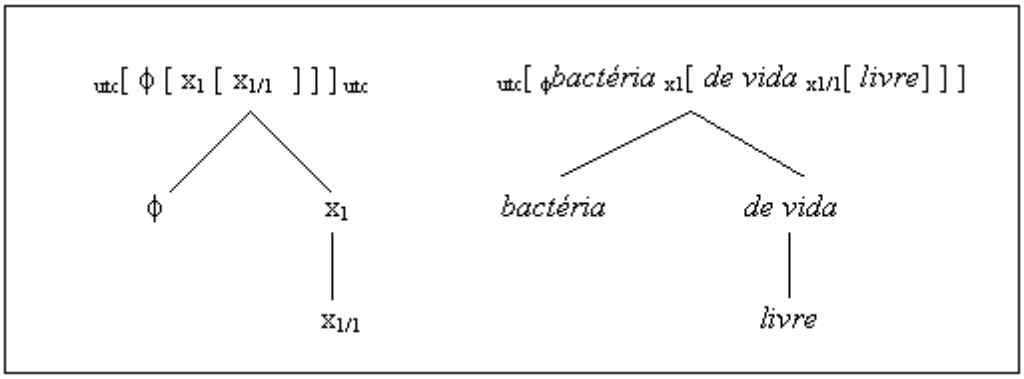
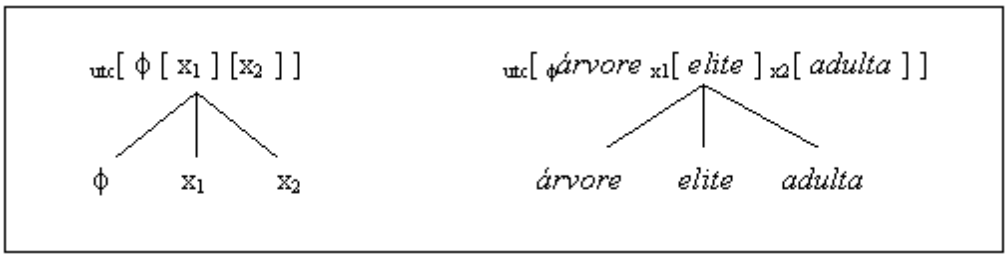


Figure 4.26 — Représentation en arbres syntagmatiques de la règle [ϕ [x_1] [x_2]]



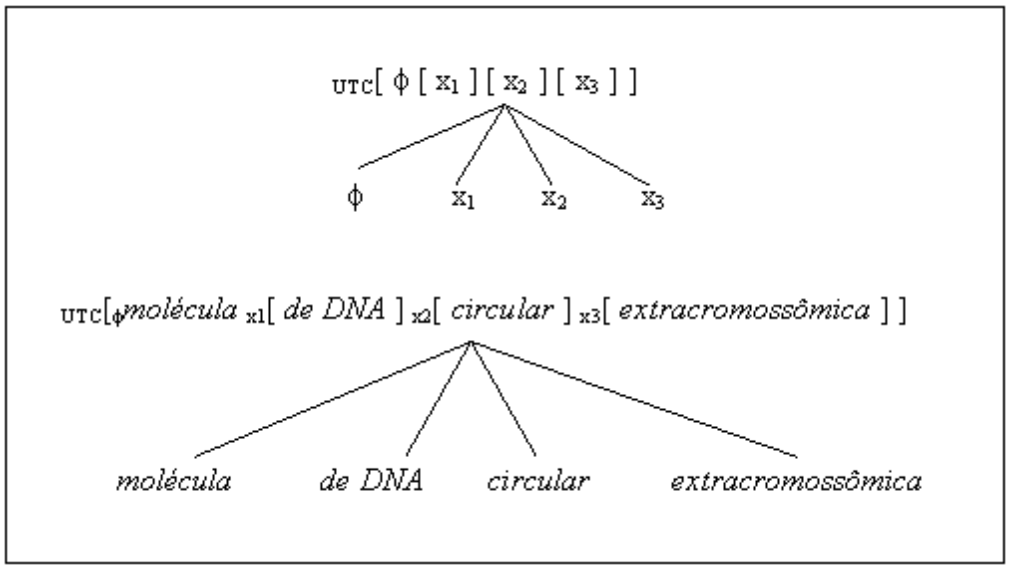
Ces deux figures font ressortir la nature différente des types de rapports entre les composants des deux UTC. L'UTC *bactéria de vida livre* (bactérie à vie autonome) est

caractérisée par un emboîtement des composants placés à droite de la base. Ainsi, la base *bactéria* (bactérie) prédique l'argument *de vida* (à vie), mais pas l'expansion *livre* (autonome). Cette dernière est prédiquée par l'argument *de vida* (à vie).

Par contre, dans le cas de l'UTC *árvore elite adulta* (arbre élite adulte), la base *árvore* (arbre) prédique à la fois l'argument *elite* (élite) et l'argument *adulta* (adulte). C'est pour cette raison que les deux composants à droite de la base sont considérés comme des arguments distincts.

Pour ce qui est des UTC de la règle [ϕ [x_1] [x_2] [x_3]], il y a un troisième argument qui a lui aussi un rapport avec la base. C'est d'ailleurs le cas de l'UTC *molécula de DNA circular extracromossômica* (molécule d'ADN circulaire extrachromosomique) présentée dans la figure suivante.

Figure 4.27 — Représentation en arbres syntagmatiques de la règle [ϕ [x_1] [x_2] [x_3]]



L'analyse de cette UTC pourrait laisser croire que les arguments *circular* (circulaire) et *extracromossômica* (extrachromosomique) font référence à *de DNA* (d'ADN) plutôt qu'à *molécula* (molécule). Toutefois, deux raisons soutiennent notre hypothèse. Premièrement,

l'argument *extracromossômica* (extrachromosomique) est au féminin, il s'accorde donc avec la base *molécula* (molécule) qui est aussi féminine et non avec *DNA* (ADN) qui est masculin. Deuxièmement, l'ADN étant une molécule, les arguments caractérisent autant la molécule que l'ADN.

4.2.6.2 Analyse morphosyntaxique

Les règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] ne présentent que très peu de différences pour ce qui est de leurs comportements syntaxique et morphologique. Nous avons regroupé dans les figures 4.28 et 4.29 les formules syntaxiques et morphologiques pour chacune de règles analysées.

Figure 4.28 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1] [x_2]]

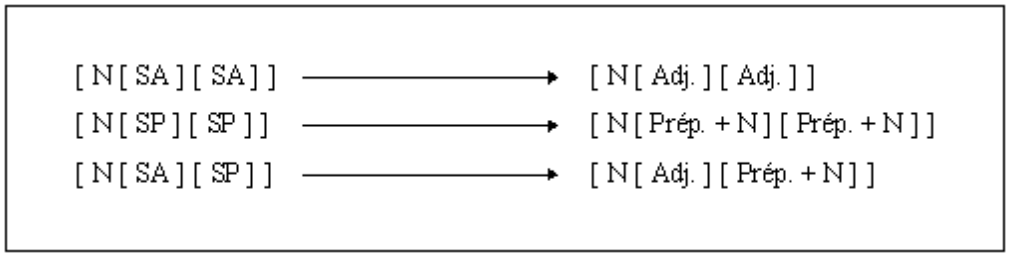
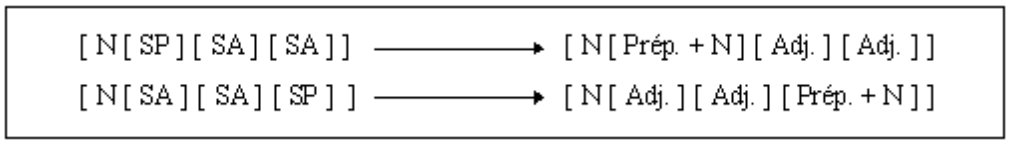


Figure 4.29 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1] [x_2] [x_3]]



Les tableaux qui suivent réunissent quelques exemples pour les formules morphologiques données pour chaque règle.

Tableau 4.27 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1] [x_2]]

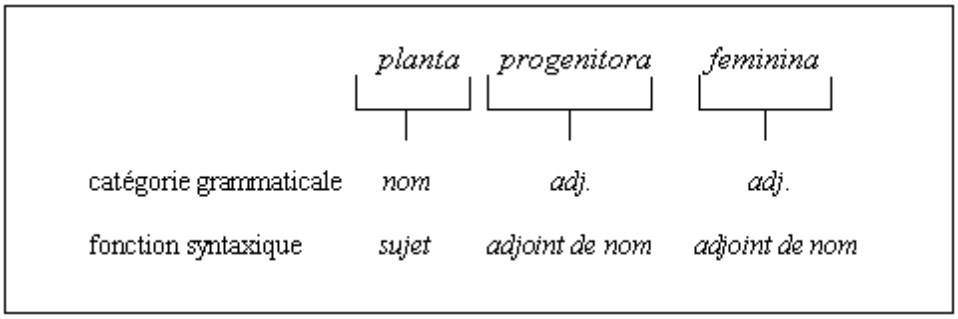
Formule morphologique	**UTC**
[N [Adj.] [Adj.]]	*linhagem parental superior* (lignée parentale supérieure)
[N [Prép. + N] [Prép. + N]]	*infecção de plantas por vírus* (infection des plantes par virus)
[N [Adj.] [Prép. + N]]	*mecanismo biológico de resistência* (mécanisme biologique de résistance)

Tableau 4.28 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1] [x_2] [x_3]]

Formule morphologique	**UTC**
[N [Prép + N] [Adj.] [Adj.]]	*molécula de DNA circular extracromossômica* (molécule d'ADN circulaire extrachromosomique)
[N [Adj.] [Adj.] [Prép. + N]]	*componente genômico nuclear de plantas* (composant génomique nucléaire de plantes)

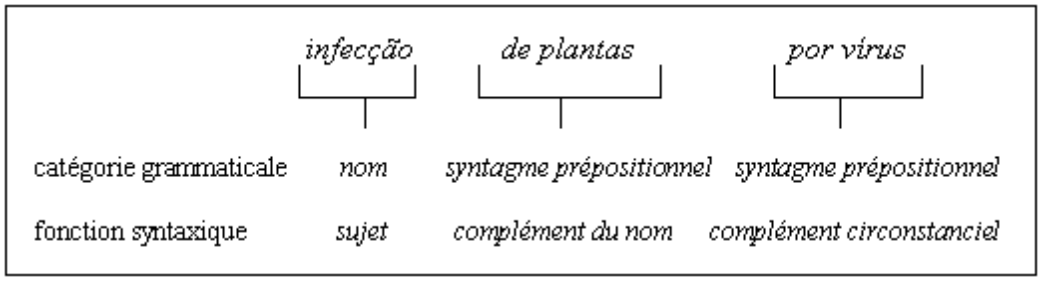
La catégorie grammaticale qui prédomine dans les arguments de la règle [ϕ [x_1] [x_2]] est l'*adjectif*. Au niveau de leur fonction syntaxique, ces arguments assument la fonction d'*adjoint de nom*. Tel est le cas de l'UTC *planta progenitora feminina* (plante *progénitrice féminine) représentée ci-dessous sous forme schématisée.

Figure 4.30 — Catégorie grammaticale et fonction syntaxique de planta progenitora feminina



Nous avons également trouvé des exemples qui ont la fonction syntaxique de *complément circonstanciel*. Dans ce cas, il ne s'agit pas d'adjectifs, mais des syntagmes prépositionnels. L'UTC *infecção de plantas por vírus* (infection des plantes par virus) en est un exemple (v. figure ci-dessous).

Figure 4.31 — Catégorie grammaticale et fonction syntaxique de infecção de plantas por vírus



Nous remarquons qu'un syntagme prépositionnel peut également assumer une fonction syntaxique de *complément du nom*. Cette fonction syntaxique est d'ailleurs la plus fréquente pour cette catégorie grammaticale. L'argument *por vírus* (par virus) est considéré comme un *complément circonstanciel* à cause de sa nature modale dans la structure de l'UTC. Autrement dit, la fonction de cet argument est de préciser par quel moyen la plante a été infectée. En effet, cet argument précise que l'infection de la plante vient d'un agent appelé virus.

4.2.6.3 Relations conceptuelles entre la base et ses arguments

Les étiquettes décrites dans le tableau ci-dessous représentent les relations conceptuelles entretenues entre la base et les arguments. Étant donné que plusieurs arguments des règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] ont des rapports avec la base, le nombre d'étiquettes varie en fonction du nombre d'arguments. Ainsi, la règle [ϕ [x_1] [x_2]], a deux étiquettes, tandis que la règle [ϕ [x_1] [x_2] [x_3]] en a trois.

Tableau 4.29 — Relations conceptuelles pour la règle [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]]

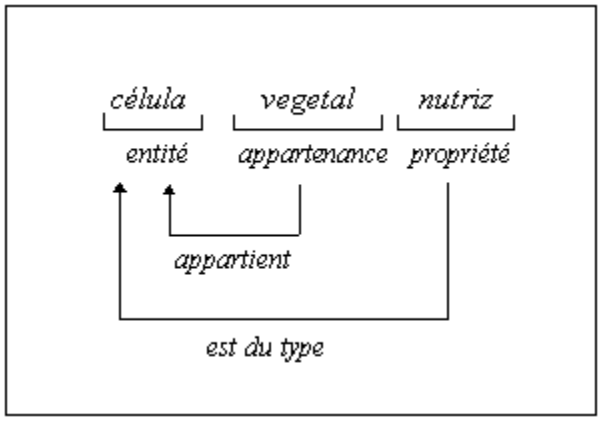
Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**			
	base	**argument 1**	**argument 2**	**argument 3**
appartient / a la forme / est du type	*entité*	*appartenance*	*forme*	*propriété*
contient / se situe dans / appartient	*entité*	*contenance*	*localisation*	*appartenance*
agit comme un / agit comme un	*agent*	*propriété*	*propriété*	
agit comme un / est du type	*agent*	*propriété*	*propriété*	
agit de manière / produit par	*processus*	*mode*	*agent*	
agit de manière / produit un effet	*processus*	*mode*	*effet*	
agit sur / a pour but de	*processus*	*patient*	*but*	
agit sur / est du type	*processus*	*patient*	*propriété*	
agit sur / produit par	*processus*	*patient*	*agent*	
appartient / est du type	*entité* *entité* *agent*	*appartenance* *appartenance* *appartenance*	*mode* *propriété* *propriété*	
appartient / est en état	*entité*	*appartenance*	*état*	
appartient / produit un effet	*processus*	*appartenance*	*effet*	
contient / se situe dans	*groupement* *entité*	*contenance* *contenance*	*localisation* *localisation*	
est du type / est du type	*groupement*	*propriété*	*propriété*	
est du type / est en état	*entité*	*propriété*	*état*	
produit un effet / est du type	*agent* *agent*	*effet* *effet*	*instrument* *propriété*	
produit un effet / est en état	*agent*	*effet*	*état*	
produit un effet / produit un effet	*agent*	*effet*	*effet*	
produit un résultat / agit de manière	*processus*	*résultat*	*mode*	
résulte de / est en état	*entité*	*source*	*état*	
se situe dans / a la forme	*entité*	*localisation*	*forme*	
se situe dans / agit comme un	*entité*	*localisation*	*propriété*	

La lecture de ce tableau montre une grande diversité des types de relations conceptuelles. Il est intéressant de constater que 22 types de relations conceptuelles se dégagent des 33 UTC de ces deux règles. Ce nombre peut sembler élevé à première vue, mais il faut préciser que ce sont en fait 14 relations qui se combinent entre elles.

Il faut également souligner qu'il y a une certaine constance d'emploi des fonctions sémantiques pour chaque ensemble de relation conceptuelle. En réalité, il n'y a que deux étiquettes qui expriment plus d'un groupe de fonctions sémantiques : *appartient / est du type* et *produit un effet / est du type*.

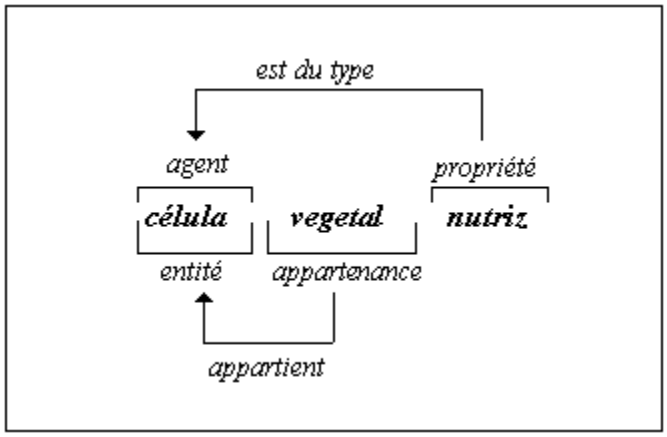
Les fonctions sémantiques peuvent parfois être interprétées de plusieurs façons. L'UTC *célula vegetal nutriz* (cellule végétale nourricière) en est un exemple. Cette UTC exprime une relation conceptuelle du type *appartient* / *est du type*. En d'autre mots, il s'agit d'une cellule qui *appartient* au domaine végétal et qui *est du type* qui nourrit. Pour exprimer cette relation conceptuelle, chaque composant de l'UTC représente une fonction sémantique. Dans le cas de l'UTC citée, la base *célula* (cellule) est une *entité*, l'argument *vegetal* (végétale) est l'*appartenance* et l'argument *nutriz* (nourricière) est une *propriété*. Nous résumons de manière schématisée ce que nous venons d'expliquer.

Figure 4.32 — Relations conceptuelles et fonctions sémantiques de célula vegetal nutriz



Étant donné que ces fonctions sémantiques sont interprétées selon la relation conceptuelle établie entre la base et plusieurs arguments, il peut parfois arriver que la base ait plus d'une fonction sémantique. Ainsi, dans cette UTC, la base *célula* (cellule) dans cette UTC pourrait également avoir la fonction d'*agent* à cause de sa *propriété* nourricière. Dans ce cas, l'UTC sera représentée de la façon suivante.

Figure 4.33 — Différentes fonctions sémantiques pour la même base d'une UTC



L'analyse que nous venons de présenter n'est qu'un exemple de la complexité d'une interprétation sémantique de la prédication. En effet, dans la structure de la prédication, la base est combinée aux arguments dans le but de fournir l'information sur un référent. Dans ce contexte, la base peut assumer plusieurs fonctions sémantiques selon le nombre d'arguments sur lesquels elle agit.

4.2.7 Cas particuliers

Dans la section qui suit, nous allons présenter trois règles de formation qui caractérisent des modes particuliers de constitution des UTC, à savoir

- $[\ \phi\ [\ x_1\ [\ x_{1/1}\ ]\ ]\ [\ x_2\ [\ x_{1/2}\ ]\ ]\ ]$;
- $[\ \phi\ [\ x_1\ ]\ [\ x_2\ [\ x_{1/2}\ [\ x_{1/1/2}\ ]\ ]\ ]\ ]$;

- [ϕ [x_1] [x_2 [$x_{1/2}$]]].

Il s'agit des UTC composées d'une *prédication nucléaire* dont les arguments s'organisent de différentes façons. Cette organisation est fondée sur le processus d'expansion de l'argument déjà cité dans d'autres règles au cours de ce chapitre. Ce qui les distinguent c'est que les règles présentement analysées peuvent spécifier non seulement le premier argument, mais aussi le deuxième. C'est précisément cet aspect que nous allons mettre en relief dans cette section.

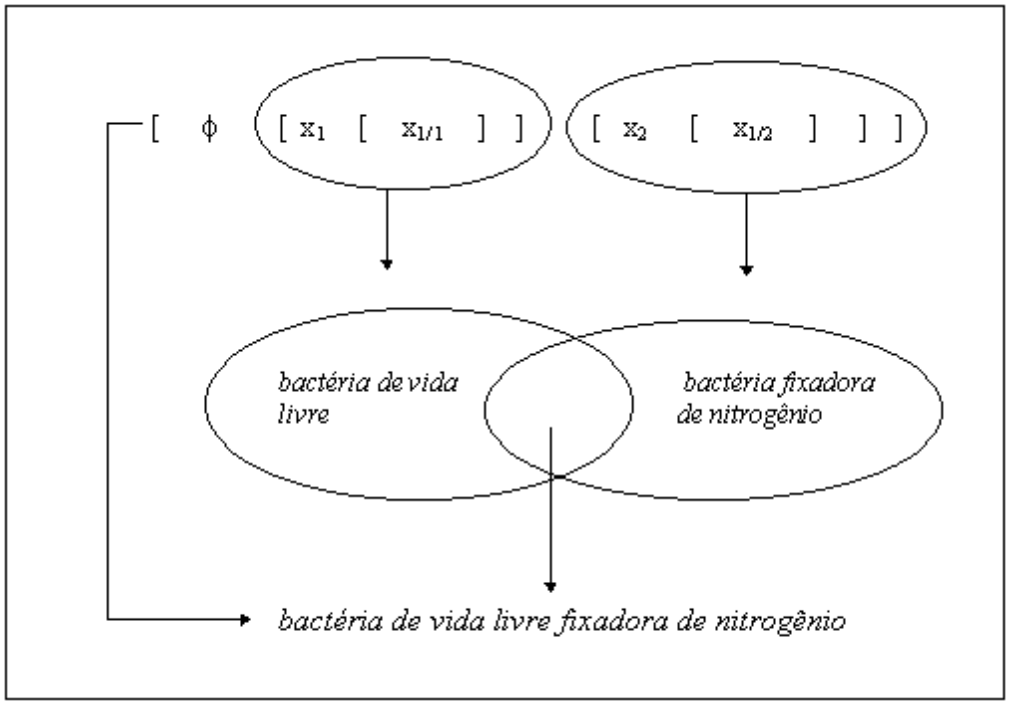
4.2.7.1 [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]]

L'UTC *bactéria de vida livre fixadora de nitrogênio* (bactérie à vie autonome fixatrice d'azote) est la seule UTC qui représente la règle de formation [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]]. Cette UTC est dérivée d'un fusionnement des UTC *bactéria de vida livre* (bactérie à vie autonome) et *bactéria fixadora de nitrogênio* (bactérie fixatrice d'azote).

Les dictionnaires et les banques de terminologie enregistrent les deux UTC de façon séparée. La banque Termium recense individuellement en français « bactérie à vie autonome » et « bactérie fixatrice d'azote ». Le glossaire multilingue *Biotechnology Glossary* enregistre pour le portugais *microrganismo fixador de nitrogênio* (micro-organisme fixateur d'azote).

Toutefois, une *bactéria de vida livre fixadora de nitrogênio* (bactérie à vie autonome fixatrice d'azote) n'est ni une simple bactérie à vie autonome ni seulement une bactérie fixatrice d'azote. Cette UTC comporte ces deux facettes simultanément. Nous illustrons dans la figure qui suit l'intersection de ces deux ensembles représentés tant par les crochets de la règle [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]] que par un graphique.

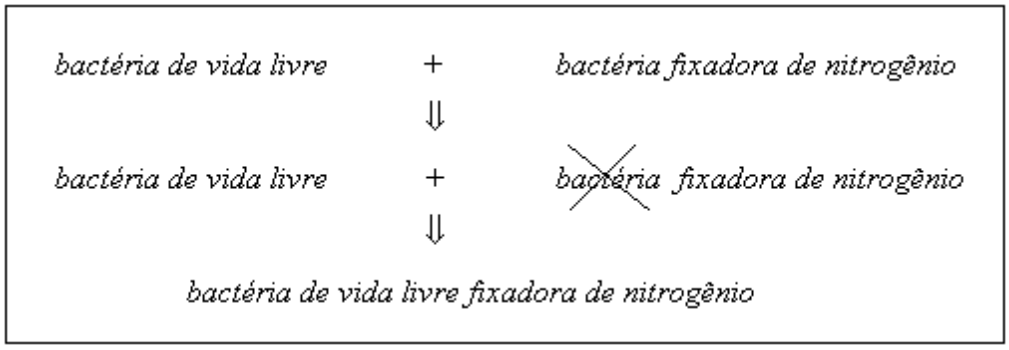
Figure 4.34 — Représentation graphique de l'UTC bactéria de vida livre fixadora de nitrogênio



Le fait d'avoir un référent qui comporte les caractéristiques des deux UTC réaffirme le caractère d'unité complexe autonome de l'UTC *bactéria de vida livre fixadora de nitrogênio* (bactérie à vie autonome fixatrice d'azote). Ainsi, la bactérie *Klebsiella pneumoniae* en est un exemple puisqu'il s'agit d'un type de bactérie qui est en même temps de vie autonome et fixatrice d'azote.

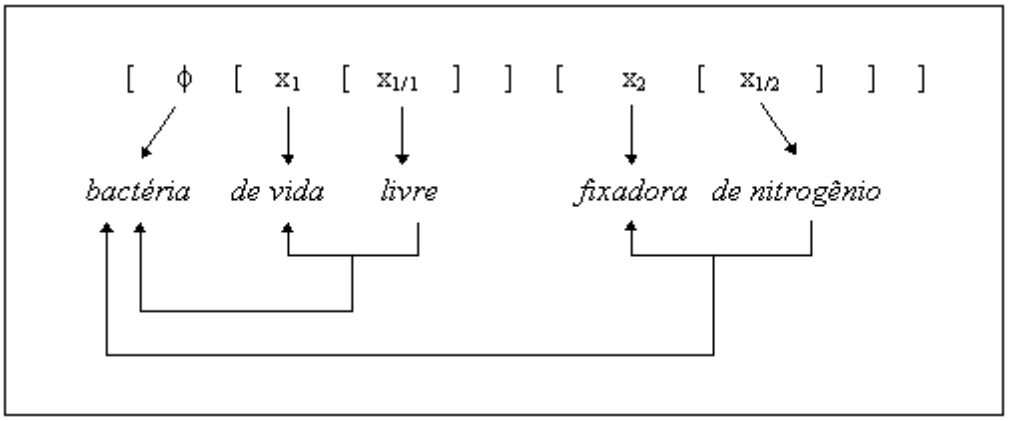
Comme le montre la figure suivante, il est possible de supposer que, pendant le processus d'union de ces deux structures, une des deux bases a été effacée.

Figure 4.35 — Processus de formation de l'UTC bactéria de vida livre fixadora de nitrogênio



La règle [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]] est formée de cinq composants. Elle comporte une base (ϕ), deux arguments (x_1 et x_2) et deux expansions d'arguments ($x_{1/1}$ et $x_{1/2}$). La première expansion constitue une spécification du premier argument, tandis que la seconde expansion établit une relation avec le second argument. Regardons ces rapports dans la figure qui suit.

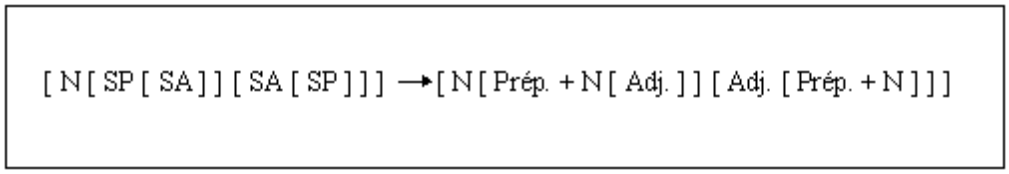
Figure 4.36 — Rapports selon la relation entretenue entre la base, les arguments et les expansions



Ainsi, l'expansion *livre* (autonome) entretient un rapport avec *vida* (vie). C'est l'ensemble *de vida livre* (à vie autonome) qui se réfère à la base *bactéria* (bactérie). Un rapport analogue est maintenu entre *de nitrogênio* (d'azote) et *fixadora* (fixatrice) et entre l'ensemble *fixadora de nitrogênio* (fixatrice d'azote) et la base *bactéria* (bactérie).

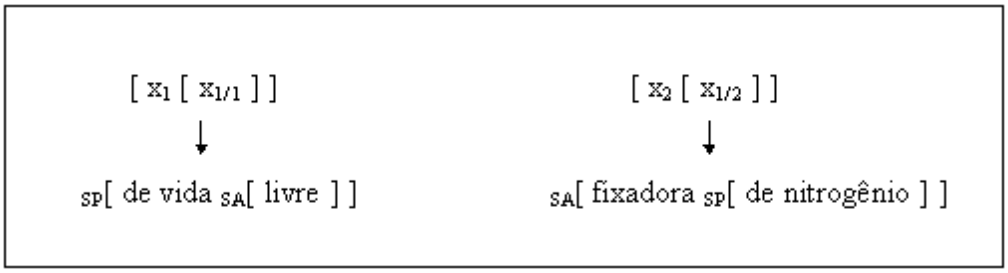
Au plan morphosyntaxique, cette UTC possède les formules syntaxique et morphologique montrées dans la figure 4.37.

Figure 4.37 — Formule syntaxique et morphologique pour la règle [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]]



La figure ci-dessous illustre l'alternance des catégories grammaticales du premier et du second groupe d'arguments suivis d'expansions. Le premier ensemble présente le patron [SP [SA]], tandis que le second résulte d'une séquence du type [SA [SP]].

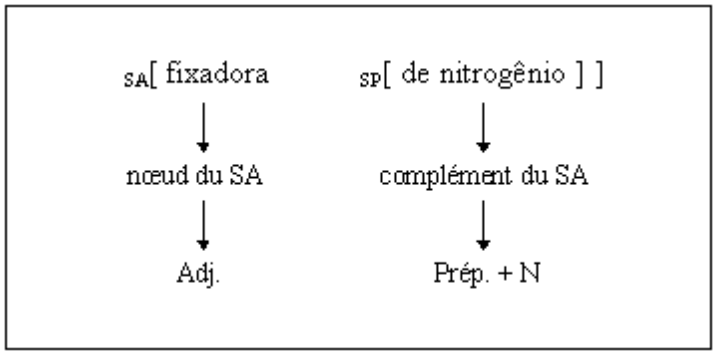
Figure 4.38 — Alternance de catégories grammaticales selon l'ensemble argument et expansion



Puisque la formule morphologique dérive de la formule syntaxique, le même phénomène se produit au plan morphologique. Il s'agit donc d'une UTC formée de deux structures morphosyntaxiques.

Pour ce qui est précisément de la séquence *fixadora de nitrogênio* (fixatrice d'azote), il s'agit d'un SA où *fixadora* (fixatrice) est le nœud du SA et *de nitrogênio* (d'azote) le complément du SA. Voici la représentation de ces catégories.

Figure 4.39 — Formation de SA dans l'UTC bactéria de vida livre fixadora de nitrogênio



Ce phénomène est décrit par Perini dans la *Gramática descritiva do português*. L'auteur mentionne que dans la structure interne du SA en portugais, le complément du SA est toujours composé d'une préposition suivie d'un SN, ou de mots formés par le suffixe *-ment*, tandis que le nœud du SA est composé de mots uniques (et non de séquences Prép. + SN), traditionnellement classés comme « adjectifs » ou « substantifs » (v. 1996 : 114).

Le cas de l'argument *fixadora* (fixatrice) s'avère intéressant. Il s'agit d'un nom dérivé du verbe *fixar* (fixer) en fonction adjectivale dans l'UTC *bactéria de vida livre fixadora de nitrogênio* (bactérie à vie autonome fixatrice d'azote).

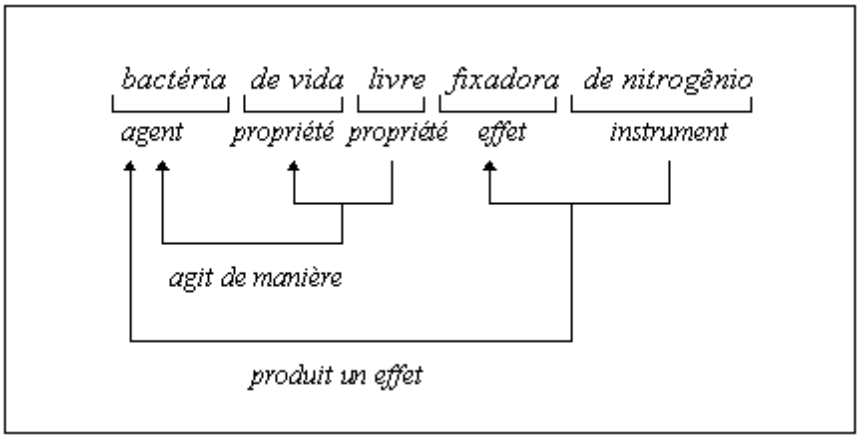
En ce qui concerne les relations conceptuelle et les fonctions sémantiques, elles sont présentées dans le tableau suivant.

Tableau 4.30 — Relations conceptuelles et fonctions sémantiques de la règle [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]]

Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**				
	base	**argument 1**	**expansion 1/1**	**argument 2**	**expansion 1/2**
agit de manière / produit un effet	*agent*	*propriété*	*propriété*	*effet*	*instrument*

Comme nous l'avons déjà mentionné, les relations conceptuelles étudiées dans cette recherche se limitent à l'analyse des rapports entre la base et les arguments. Seules les fonctions sémantiques sont interprétées au niveau de chaque composant de l'UTC. Ainsi, la base-*agent* (*bactéria*) *agit de la manière* exprimée par les *propriétés* de l'argument 1 plus son expansion (*de vida livre*) et produit un *effet* (*fixadora*) exprimé par l'argument 2 et spécifié par son expansion-*instrument* (*de nitrogênio*). Les fonctions sémantiques des expansions viennent compléter la notion des arguments. L'expansion 1/2, par exemple, est l'*instrument* de l'*effet* produit par l'argument 2. Dans la figure 4.40, l'UTC *bactéria de vida livre fixadora de nitrogênio* (bactérie à vie autonome fixatrice d'azote) illustre ces rapports.

Figure 4.40 — Relations conceptuelles et fonctions sémantiques de bactéria de vida livre fixadora de nitrogênio

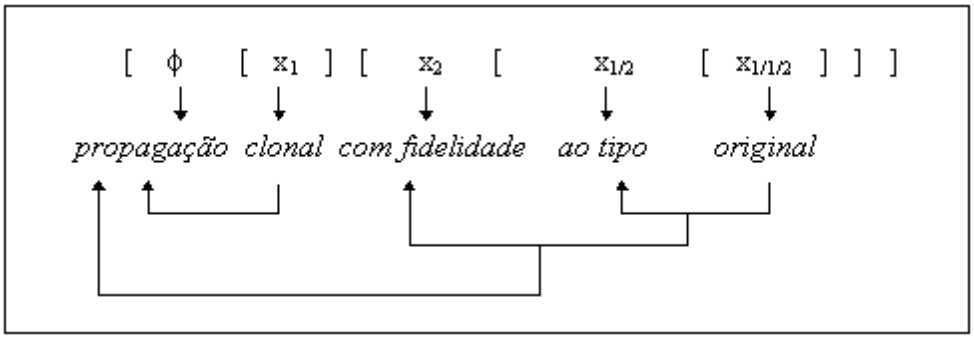


Cette figure montre que la bactérie est un agent (*bactéria*) qui agit de manière libre (*de vida libre*) et produit un effet qui consiste à fixer l'azote (*fixadora de nitrogênio*).

4.2.7.2 [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]]

Cette règle diffère de la précédente par le fait qu'elle contient deux niveaux d'expansion à droite du second argument. Comme la règle antérieure, cette règle a été relevée une seule fois dans le corpus. Nous avons retenu l'UTC qui la représente étant donné l'existence des équivalents en langue anglaise et française. Il s'agit de l'UTC *propagação clonal com fidelidade ao tipo original* dont les équivalents anglais et français sont respectivement « true-to-type clonal propagation » et « propagation clonale stable ». La figure suivante reprend cette UTC et montre la façon dont nous avons analysé l'enchaînement des composants situés à droite de la base.

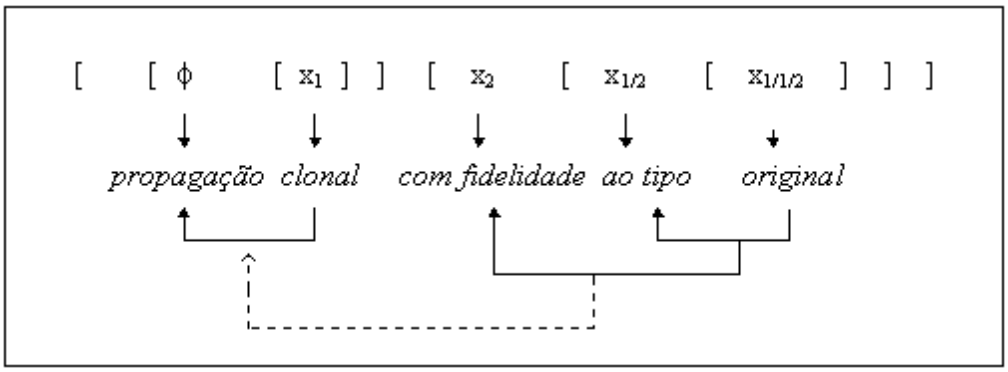
Figure 4.41 — Rapports selon la relation entre la base, les arguments et les expansions pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]]



Cette UTC s'analyse en deux groupes. Le premier groupe se réfère à la relation entretenue entre la séquence *com fidelidade ao tipo original* et la base *propagação*. Ainsi, l'expansion $x_{1/1/2}$ représentée par le composant *original* a une incidence sur l'expansion $x_{1/2}$ (*ao tipo*). L'ensemble *ao tipo original* entretient un rapport avec le second argument *com fidelidade*. Enfin, toute la séquence *com fidelidade ao tipo original* a un rapport avec la base *propagação*. Le second groupe fait référence au rapport maintenu entre l'argument *clonal* (clonale) et la base *propagação* (propagation).

Nous pouvons également analyser le rapport entretenu par la paraphrase *com fidelidade ao tipo original* comme faisant référence au groupe *propagação clonal* (propagation clonale) puisqu'il ne s'agit pas uniquement de produire n'importe quoi qui serait fidèle au type original, mais des clones. Cette interprétation résulte en une autre représentation de la règle comme le montre l'illustration qui suit.

Figure 4.42 — Nouvelle représentation des rapports entre les composants de l'UTC propagação clonal com fidelidade ao tipo original



Il faut souligner que cette nouvelle analyse implique des changements dans la façon d'établir la règle de formation. Comme on peut le voir, de nouveaux crochets ont été insérés pour bien démarquer la séquence *propagação clonal* (propagation clonale) et son rapport avec le groupe *com fidelidade ao tipo original*.

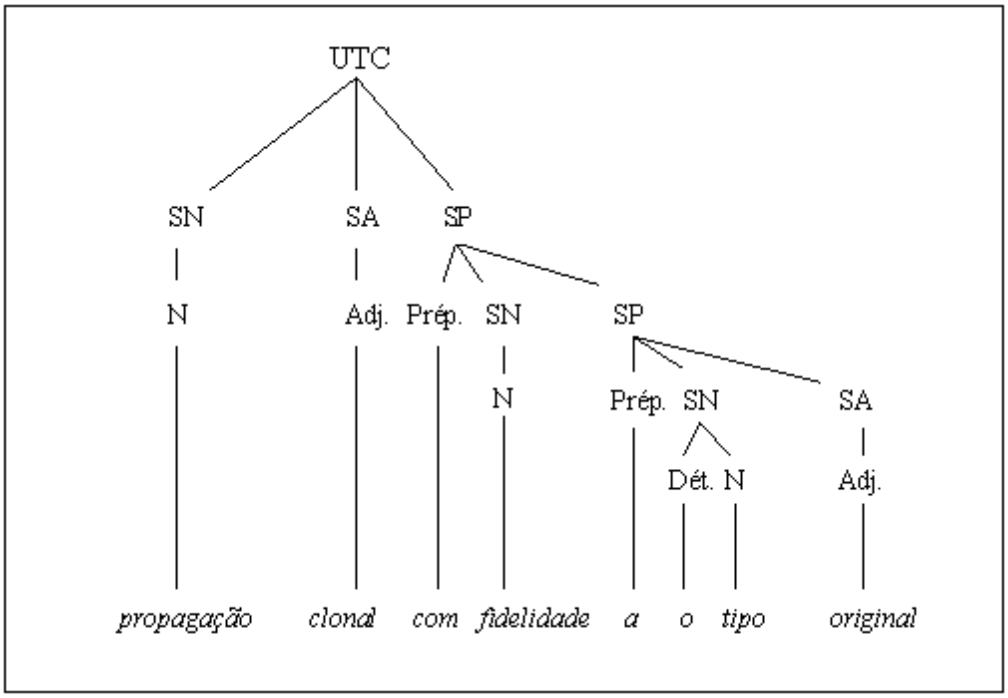
Dans le cadre de l'analyse morphosyntaxique, cette UTC caractérise les formules décrites dans la figure qui suit.

Figure 4.43 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]]

[N [SA] [SP [SP [SA]]]] → [N [Adj.] [Prép. + N [Prép. + Dét. + N [Adj.]]]]

L'analyse en forme d'arbre syntagmatique (v. figure suivante) aide à comprendre le chemin parcouru pour passer de la formule syntaxique à la formule morphologique.

Figure 4.44 — Arbre syntagmatique de propagação clonal com fidelidade ao tipo original



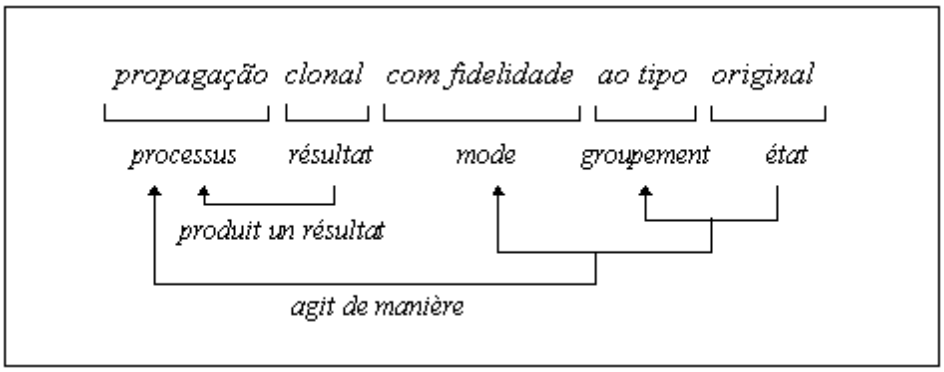
Les rapports morphosyntaxiques établis à l'intérieur de la prédication nucléaire sont fondés sur une analyse des rôles sémantiques joués par chaque composant de l'UTC. À ce sujet, nous présentons dans le prochain tableau les relations conceptuelles qui sont le fruit des rapports sémantiques entre les composants de l'UTC *propagação clonal com fidelidade ao tipo original* (propagation clonale stable).

Tableau 4.31 — Relations conceptuelles pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]]]

Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**				
	base	**argument 1**	**argument 2**	**expansion 1/2**	**expansion 1/1/2**
produit un résultat / agit de manière	*processus*	*résultat*	*mode*	*groupement*	*état*

Ainsi, l'UTC *propagação clonal com fidelidade ao tipo original* (propagation clonale stable) est une propagation (*processus*) qui produit comme résultat des clones (*résultat*) et agit de façon à ce qu'ils restent fidèles (exacts) au type original (*mode* + *groupement* + *état*). L'interprétation des fonctions sémantiques du second argument et de ses expansions est complexe comme le montre la figure ci-dessous.

Figure 4.45 — Relations conceptuelles et fonctions sémantiques de com fidelidade ao tipo original



La paraphrase *com fidelidade ao tipo original* traduite en français par « stable » et en anglais par « true-to-type » signifie que les gènes produits par propagation doivent résulter fidèlement (*mode*), comme le type (*groupement*) des gènes, d'un *état* original.

4.2.7.3 [ϕ [x_1] [x_2 [$x_{1/2}$]]]

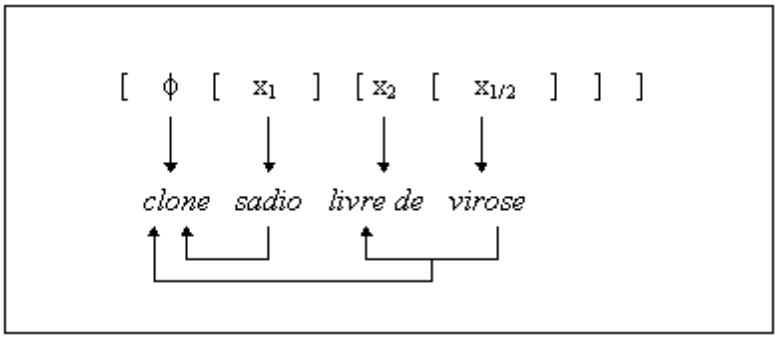
La règle [ϕ [x_1] [x_2 [$x_{1/2}$]]] se caractérise par une expansion du second argument. Quatre UTC recueillies dans notre corpus sont formées selon cette règle. Ces UTC emploient des adverbes qui agissent comme arguments ou des prédicats qui marquent l'absence du référent exprimé par l'expansion de l'argument. Voici les exemples recensés.

- *célula interna altamente vacuolizada* (cellule interne hautement vacuolaire);
- *célula somática geneticamente diferente* (cellule somatique génétiquement différente);
- *clone sadio livre de virose* (clone sain exempt de virose);
- *tecido somático geneticamente distinto* (tissu somatique génétiquement distinct).

Il est intéressant de noter que, dans le domaine de la biotechnologie des végétaux, l'argument *livre de* (exempt de) dans l'UTC *clone sadio livre de virose* (clone sain exempt de virose) est fréquemment utilisé. Comme le souligne Faulstich, ce prédicat peut alterner avec d'autre prédicats comme *isento de* (exempt de), ou par l'ajout du préfixe *a-* qui marque lui aussi l'absence (v. s. d. b : 9).

La représentation de l'UTC *clone sadio livre de virose* (clone sain exempt de virose) est illustrée dans la prochaine figure.

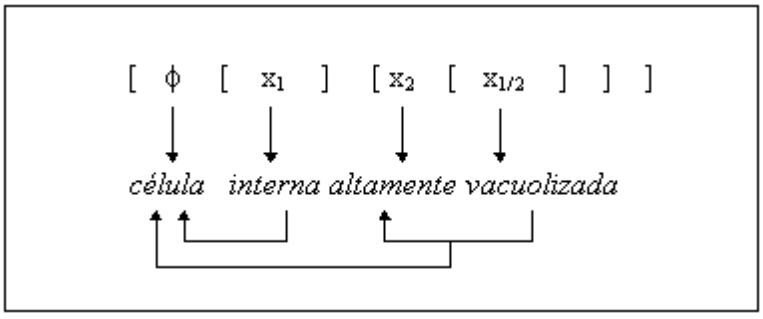
Figure 4.46 — Rapports selon la relation entre la base, les arguments et l'expansion pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]]



L'expansion *virose* (virose) entretient un rapport avec l'argument *livre de* (exempt de). L'ensemble *livre de virose* (exempt de virose) se réfère à la base *clone* (clone). Tout comme d'autres cas analysés auparavant, nous pouvons considérer que la séquence *livre de virose* (exempt de virose) maintient un rapport avec *clone sadio* (clone sain) pris comme un tout.

Pour ce qui est des UTC composées d'un argument-adverbe, par exemple *célula interna altamente vacuolizada* (cellule interne hautement vacuolaire), leurs rapports se font de la façon décrite dans la figure suivante.

Figure 4.47 — Rapports selon la relation entre la base, l'argument, l'argument-adverbe et l'expansion pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]]



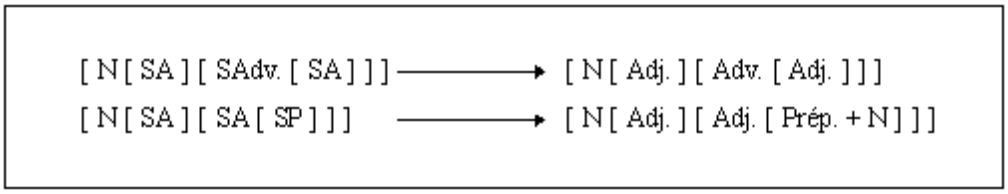
Comme nous pouvons le remarquer, il y a deux ensembles de relations. Le premier, plus simple, contient le rapport entre l'argument *interna* (interne) et la base *célula* (cellule). Le second représente le rapport entretenu entre la séquence *altamente vacuolizada* (hautement vacuolaire) et la base *célula* (cellule). Dans ce dernier cas, il faut signaler le rapport entre l'adverbe *altamente* (hautement) et l'adjectif *vacuolizada* (vacuolaire). En effet, cette structure adverbe + adjectif a déjà été constatée dans les règles [ϕ [x_1 [$x_{1/1}$]]] et [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]. Comme nous l'avons fait pour les autres règles, la structure présentement analysée est interprétée au plan fonctionnel comme un *adverbe nucléaire*. Cette dénomination vient de la position que l'adverbe occupe dans la structure de la prédication. Les *adverbes nucléaires* font partie de la prédication nucléaire puisqu'ils sont considérés comme des arguments. Au niveau de sa nature, l'adverbe *altamente* (hautement) appartient à la langue générale, parce qu'il est composé de l'adjectif *haut* qui n'appartient pas en propre aux langues de spécialité. Toutefois, nous avons également trouvé des UTC composées d'adverbes formés à partir d'adjectifs du domaine suivis par le suffixe *-mente*. Nous les appelons *adverbes de domaine* conformément à Guimier (v. supra, p. 36). Voici les exemples qui figurent dans notre corpus.

- *célula somática* ***geneticamente*** *diferente* (cellule somatique génétiquement différente);
- *tecido somático* ***geneticamente*** *distinto* (tissu somatique génétiquement distinct).

Au niveau syntaxique, les adverbes sont considérés comme des compléments circonstanciels. Les adjectifs qui les suivent sont traités comme des adjoints de nom. À notre avis, c'est la séquence adverbe plus adjectif qui précise le signifié de la base et non chaque composant pris séparément.

Pour ce qui est des structures morphosyntaxiques, la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]] comporte deux types de formules syntaxiques qui se caractérisent de deux manières différentes au niveau morphologique et qui sont représentées dans la figure ci-dessous.

Figure 4.48 — Formules syntaxiques et morphologiques pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]]



Les formules morphologiques sont présentées dans le tableau qui suit avec les exemples recensés dans notre corpus.

Tableau 4.32 — Exemples d'UTC par formule morphologique pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]]

Formule morphologique	**UTC**
[N [Adj.] [Adj. [Prép. + N]]]	*clone sadio livre de virose* (clone sain exempt de virose)
[N [Adj.] [Adv. [Adj.]]]	*célula somática* ***geneticamente*** *diferente* (cellule somatique génétiquement différente);

Quant aux relations conceptuelles établies par les UTC qui caractérisent cette règle, nous en avons relevé trois ensembles. Ces relations varient selon les rôles sémantiques joués par chaque composant de l'UTC. Dans le prochain tableau nous systématisons les relations conceptuelles entretenues entre la base et les arguments, ainsi que les fonctions sémantiques jouées par chaque composant de l'UTC.

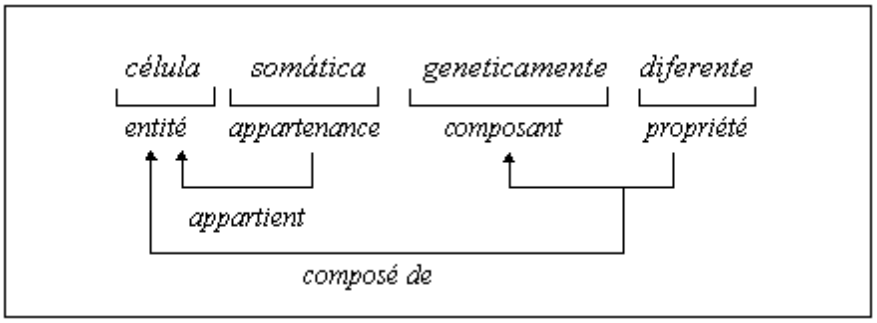
Tableau 4.33 — Relations conceptuelles pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]]

Relation conceptuelle	**Fonction sémantique de chaque composant de l'UTC**			
	base	**argument 1**	**argument 2**	**expansion 1/2**
appartient / composé de	*entité*	*appartenance*	*composant*	*propriété*
est en état /contient	*entité*	*état*	*contenance*	*contenance*

se situe dans / contient	*entité*	*localisation*	*intensité*	*contenance*

La relation *appartient / composé de* est représentée par les UTC *célula somática* ***geneticamente*** *diferente* (cellule somatique génétiquement différente) et *tecido somático* ***geneticamente*** *distinto* (tissu somatique génétiquement distinct). Nous avons repris la première UTC pour illustrer notre analyse sémantico-conceptuelle (v. figure suivante).

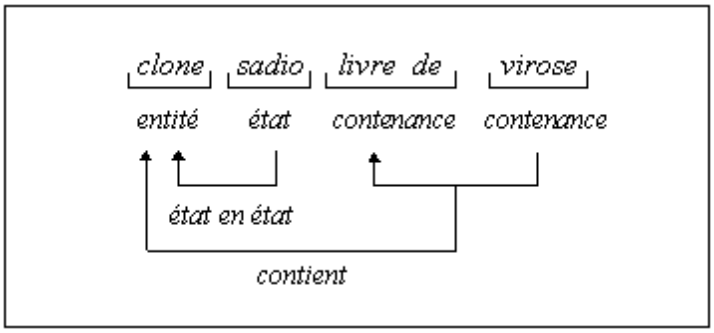
Figure 4.49 — Relations conceptuelles et fonctions sémantiques de célula somática geneticamente diferente



Ainsi, l'expansion *diferente* (différente) est une *propriété* du *composant geneticamente* (génétiquement). L'ensemble *geneticamente differente* (génétiquement différente) entretient un rapport avec la base *célula* (cellule). Autrement dit, la cellule est *composée de* gènes qui sont différents. L'argument *somática* (somatique) insère une relation d'*appartenance* avec la base *célula* (cellule) puisqu'il indique que la cellule *appartient* au soma, au corps.

La relation *est en état / contient* caractérise l'UTC *clone sadio livre de virose* (clone sain exempt de virose). Cette UTC est illustrée dans la prochaine figure avec ses relations conceptuelles et ses fonctions sémantiques.

Figure 4.50 — Relations conceptuelles et fonctions sémantiques de clone sadio livre de virose



Comme le montre la figure précédente, *virose* (virose) joue le rôle de *contenance* par rapport à l'argument *livre de* (exempt de). En réalité, l'argument *livre de* (exempt de) exprime plutôt la non-contenance ou l'absence. Nous avons attribué à l'argument *livre de* (exempt de) la fonction sémantique de *contenance*, puisque, en effet, cette expression introduit la notion de contenance zéro qui est la non-contenance et non pas l'absence en soi. Cette constatation prouve la dépendance de l'argument *livre de* (exempt de) avec son expansion *virose* (virose) qui, à son tour, a la fonction sémantique d'informer sur la non-contenance de l'argument. La séquence *livre de virose* (exempt de virose) entretient donc un rapport avec l'entité *clone* (clone) pour préciser ce qui ne contient pas de virose.

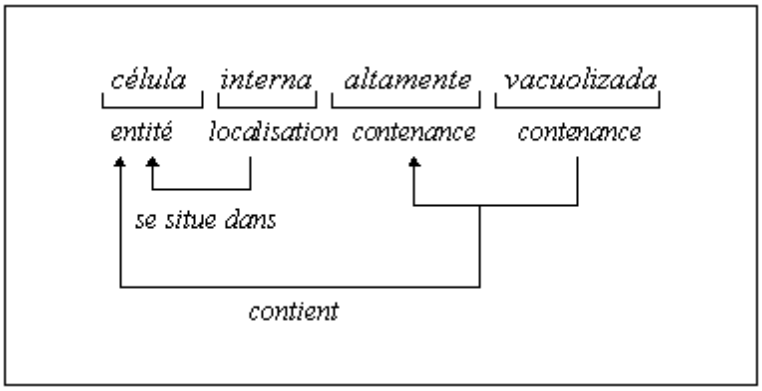
La relation *est en état / contient* représente également le rapport entre l'état sain exprimé par l'argument *sadio* (sain). Autrement dit, cet argument a la fonction sémantique d'informer sur le fait que l'entité *clone* (clone) *est en état* sain.

Fait intéressant à noter, les deux arguments *sadio* (sain) et *livre de virose* (exempt de virose) constituent différentes façons d'exprimer que le clone n'est pas malade, donc en bonne santé. Un argument renforce l'autre en utilisant divers moyens.

Finalement, la dernière relation (*se situe dans / contient*) caractérise l'UTC *célula interna altamente vacuolizada* (cellule interne hautement vacuolaire). Nous la montrons en

faisant ressortir ses fonctions sémantiques et ses relations conceptuelles dans la figure qui suit.

Figure 4.51 — Relations conceptuelles et fonctions sémantiques de célula interna altamente vacuolizada



L'UTC *célula interna altamente vacuolizada* (cellule interne hautement vacuolaire) maintient deux ensembles de relations conceptuelles. La première relation conceptuelle (*se situe dans*) exprime le rapport entre l'argument locatif *interna* (interne) et la base entité *célula* (cellule). Autrement dit, l'argument précise que la cellule *se situe* à l'intérieur. La seconde relation (*contient*) exprime le rapport entre la *contenance*, qui est *intense*, et l'*entité* qui la contient, c'est-à-dire que la cellule (*entité*) contient une quantité élevée (*intensité*) de vacuoles[58] (*contenance*). Ainsi, l'ensemble *se situe dans / contient* renferme l'idée d'une cellule située à l'intérieur qui contient beaucoup de vacuoles.

4.3 Conclusion

Il a été question au cours du présent chapitre des dix règles de formation des UTC relevées dans notre corpus. En adoptant une perspective d'analyse fonctionnaliste de la langue, nous avons mis en évidence divers aspects morphologiques, syntaxiques et sémantiques de chaque ensemble terminologique. Toutes les règles ont d'abord été

examinées en profondeur en prenant comme base les mêmes paramètres d'analyse, puis nous nous sommes penchée principalement sur les aspects qui les distinguent.

L'analyse des UTC formées d'après la règle [ϕ [x_1]], a fait ressortir que la grande majorité des *bases* présentes dans ce type de formation sont *statiques*. Cette caractéristique, renforcée par la fréquence élevée de ce type de structure, démontre la forte tendance de l'emploi nominatif dans le discours terminologique.

Nous avons également vu que, sur le plan de la syntaxe, cette règle se caractérise essentiellement pas le patron [N [SA]]. L'une des caractéristiques intéressantes de ce patron est de regrouper des UTC formées du préfixe de négation *não*. Le corpus compte aussi des UTC composées selon le modèle [N [SP]]. Il ressort de cette formule syntaxique que certaines UTC comportent un article à l'intérieur de leur structure morphologique. Nous avons également noté que la préposition contenue dans la structure [N [SP]] a une valeur de composant qui marque des rapports significatifs dans la prédication nucléaire.

Les divers composants de la règle [ϕ [x_1]] jouent plusieurs rôles syntaxiques. Ainsi, la *base* constitue le *sujet* de l'UTC, tandis que les *arguments* peuvent occuper la place d'*adjoint de nom*, de *complément du nom* ou de *complément circonstanciel*. D'ailleurs, la fonction syntaxique de la *base* est la même pour toutes les UTC puisque, comme nous l'avons affirmé à plusieurs reprises, notre corpus est composé uniquement d'UTC de base nominale.

Toujours en ce qui concerne la règle[ϕ [x_1]], nos données ont fait ressortir qu'il existe treize types de relations conceptuelles qui sont le fruit des fonctions sémantiques des bases et des arguments de l'UTC. Au niveau de l'emploi de ces fonctions sémantiques, la

[58] Vacuole : espace circonscrit, parfois limité par une membrane, au sein du cytoplasme d'une cellule ou d'un organisme unicellulaire, à contenu variable (v. Le Nouveau Petit Robert, 1995).

fonction *entité* est hautement utilisée par les *bases* (41 %) tandis qu'elle n'est employée que par 5 % des arguments.

Pour ce qui est des règles [ϕ [y_1]] et [[ϕ [x_1]] [y_1]], c'est la présence des satellites qui les distinguent de tous les autres modes de formation. Les satellites caractérisent une *prédication étendue*, autrement dit une prédication qui est en rapport avec la *prédication nucléaire* prise comme un tout et non seulement avec la base comme c'est le cas des arguments. Nous avons par contre remarqué que c'est uniquement dans la règle [ϕ [y_1]] que les satellites entretiennent un rapport avec la base en raison d'un effacement de l'argument au cours du processus de formation de l'UTC.

Les satellites sont, au niveau grammatical, soit des syntagmes prépositionnels, soit des locutions adverbiales. Dans notre corpus, ces locutions adverbiales sont les expressions latines : *in vivo, in vitro* et *in situ*. Au plan syntaxique, ces satellites sont des compléments circonstanciels et, au niveau fonctionnel, ils sont classés comme des locutions adverbiales étendues. Cette dernière étiquette prend en considération la place occupée par cet élément dans la structure de la prédication. D'un point de vue conceptuel, les satellites comportent deux types de fonctions sémantiques : *localisation* et *état*. Ils entretiennent ainsi des rapports avec les autres composants de l'UTC dans le but de représenter des relations conceptuelles.

La présence de la catégorie grammaticale locution adverbiale a donné lieu à de nouveaux patrons syntaxiques pour les règles [[ϕ [x_1]] [y_1]] et [ϕ [y_1]]. En plus de la formule [[N [SP]] [SP]], la règle [[ϕ [x_1]] [y_1]] présente les patrons [[N [SA]] [SAdv.]] et [[N [SP]] [SAdv.]]. Pour ce qui est de la règle [[ϕ [y_1]], seule la formule [N [SAdv.]] caractérise le patron syntaxique de ce mode de formation.

Les règles [ϕ [x_1 [$x_{1/1}$]]] et [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] ont pour caractéristique principale l'expansion du premier argument. Dans le cas de la règle [ϕ [x_1 [$x_{1/1}$]]], l'argument est modifié une seule fois, tandis que dans le cas de la règle [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$

]]]], non seulement l'argument est modifié, mais son expansion l'est également. Ce phénomène est interprété dans notre recherche comme un processus d'emboîtement d'UTC. En effet, il s'agit de l'insertion d'une UTC dans la structure de l'UTC principale. Il s'agit d'un recours utilisé pour former un concept donné à partir des concepts et des structures terminologiques qui existent déjà. La présence d'arguments-adverbes et d'expansions-adverbes a été vérifiée dans ces deux types de règles de formation. Ces adverbes sont formés à partir d'un adjectif auquel on ajoute le préfixe *-mente*. Au plan fonctionnel, ils sont interprétés comme des adverbes nucléaires puisqu'ils occupent une place dans la prédication nucléaire. Dans le cas spécifique de la règle [ϕ [x_1 [$x_{1/1}$]]], ces adverbes peuvent être formés à partir d'un adjectif du domaine ou de la langue générale. Au niveau des relations conceptuelles, le rapport sémantique de l'adverbe est établi avec l'adjectif qui le suit. C'est d'ailleurs la séquence adverbe + adjectif qui spécifie ou modifie la base de l'UTC.

Les règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]] ont, respectivement, le même nombre de composants que les règles [ϕ [x_1 [$x_{1/1}$]]] et [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]. La différence entre ces deux paires de règles se situe au niveau des rapports établis entre les composants. Ainsi, dans le cas des règles [ϕ [x_1] [x_2]] et [ϕ [x_1] [x_2] [x_3]], il n'existe pas de rapports entre les arguments. Ces derniers entretiennent plutôt un lien direct avec la base de l'UTC.

Finalement, les règles [ϕ [x_1] [x_2 [$x_{1/2}$]]], [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]]] et [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]] ont été considérées comme des cas particuliers étant donné la complexité de leurs structures. Ces modes de formation sont caractérisés par trois types d'expansions : une expansion du second argument pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$]]], une expansion de l'expansion du second argument pour la règle [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]]] et une expansion du premier et du second arguments pour la règle [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]]. Même si ces UTC sont peu nombreuses, elles permettent de réaliser des interprétations importantes aux niveaux morphologiques, syntaxiques et sémantiques.

5. Stratégies pour l'automatisation de l'extraction des UTC

5.1 Introduction

Jusqu'à maintenant, l'étude des UTC du domaine de la biotechnologie des végétaux nous a permis d'examiner différents aspects morphologiques, syntaxiques et sémantiques sur la base d'une analyse fonctionnaliste de la langue. Les critères que nous allons présenter dans ce chapitre sont fondés sur les résultats que nous avons obtenus au cours de l'analyse et de l'interprétation des données. Ces critères s'insèrent dans une vision plus globale de la conception d'un logiciel d'automatisation et ils ne prétendent pas résoudre toute la problématique liée à cette conception. Les idées qui seront exploitées dans les pages qui suivent présupposent que certaines tâches sont déjà accomplies par d'autres logiciels. Nous y reviendrons au fur et à mesure de l'explication.

En plus de proposer des pistes pour l'élaboration d'un logiciel d'extraction d'UTC en contexte, ce chapitre a pour but de contribuer à la conception d'instruments automatiques d'aide au traitement de problèmes linguistiques liés à la terminologie. Ainsi, nous tenterons de formaliser les règles de la structure de la prédication terminologique proposées dans notre étude pour qu'elles puissent constituer un fonds de connaissance linguistique utile pour le processus d'automatisation.

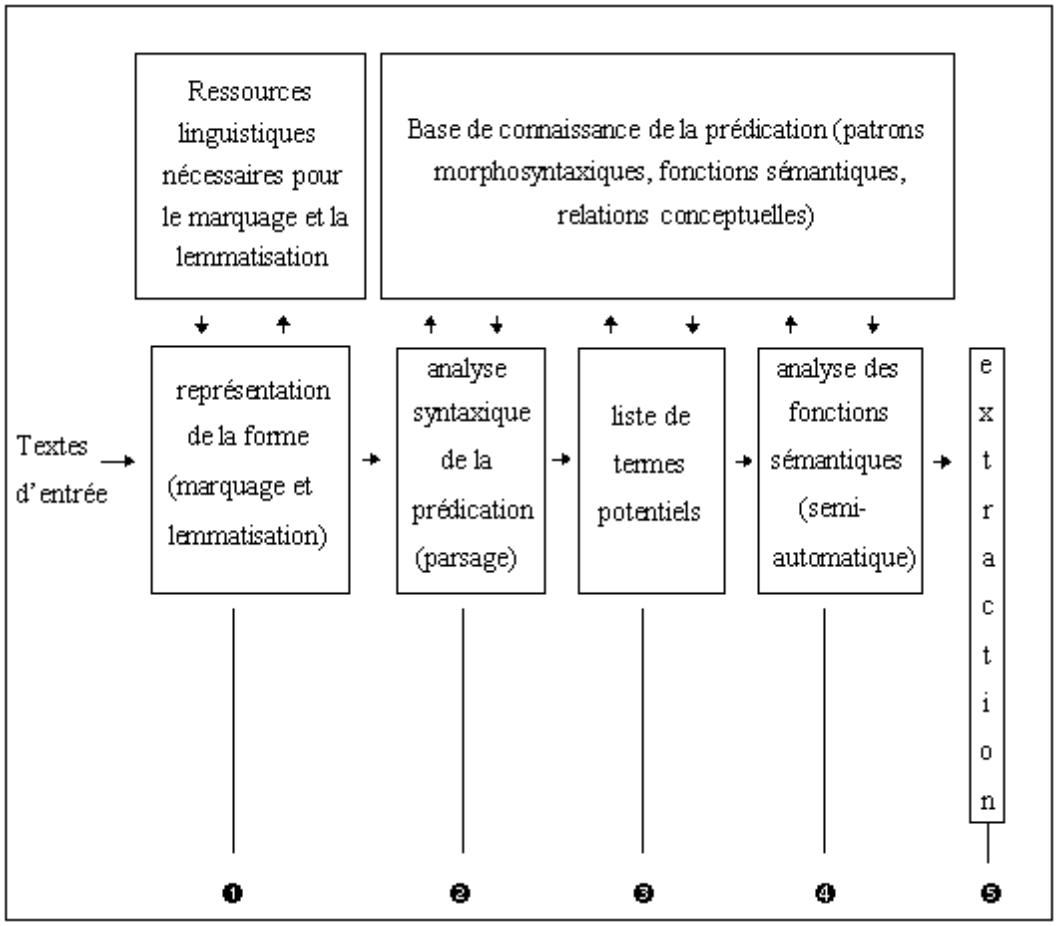
5.1.1 Considérations de base

Parmi les divers logiciels d'extraction d'UTC qui existent, nous croyons que ceux qui donnent les résultats les plus satisfaisants pour un travail linguistique sont ceux qui tirent profit de l'intelligence artificielle et qui sont orientés vers la modélisation de connaissances linguistiques. Tel est le cas des logiciels programmés en Prolog. La théorie de la grammaire fonctionnelle, par exemple, a déjà été traitée au moyen de ce type de programmation. Dans son article sur la simulation de la grammaire fonctionnelle en Prolog, Samuelsdorff justifie le choix d'un tel type de langage de programmation en mentionnant que « [...] its database contains only two kinds of items: facts and rules; these correspond to lexicon and grammar in a language theory. Further, the format of the facts in PROLOG is in terms of logical predications. This makes it possible to code the lexical entries in the form of predicates with variables in the argument position where the program may insert the terms selected by the user » (1989 : 29 et 30). En plus, Dik signale que « There is a certain convergence on PROLOG as a language in which the FG predications are most easily represented. This is understandable, since both FG and PROLOG have been inspired, to different degrees, by principles of predicate logic, and since PROLOG is a flexible tool for representing and processing symbolic relational information of the kind which is also central to FG [...] » (1989a : 10).

Il existe également d'autres expériences qui, programmées en Prolog, ont donné lieu à des applications intéressantes dans le domaine de l'élaboration des outils d'aide au traitement automatique de langues. Tel est le cas du logiciel ZStation (v. *supra* p. 73) qui utilise une interface « programmée en PDC-Visual Prolog sur la plate-forme Windows » (Zinglé 1996 : 79).

En prenant en considération les aspects mentionnés ci-dessus sur l'approche idéale en matière d'informatique et la capacité du logiciel Prolog de les prendre en compte, nous suggérons, pour commencer, une structure générale de l'environnement du traitement terminologique. Cette structure est proposée dans la prochaine figure.

Figure 5.1 — Structure générale du traitement automatique pour l'extraction des UTC



L'environnement de traitement terminologique proposé dans cette figure est divisé en cinq étapes. Examinons chaque étape séparément. La première étape concerne le traitement grammatical des textes d'entrée. Il s'agit de la représentation de la forme des mots contenus dans les textes écrits. Ainsi, ces textes, déjà numérisés, sont analysés par un programme de marquage de mots qui étiquette les catégories grammaticales et par un logiciel de lemmatisation pour l'attribution des lemmes. La lemmatisation, en particulier, doit tenir compte des problèmes reliés aux formes ambiguës. Ces processus de marquage et de lemmatisation doivent recourir à diverses ressources linguistiques, par exemple les grammaires flexionnelles ainsi que d'autres dictionnaires qui seront alimentés au fur et à mesure que de nouvelles données s'ajoutent au système. De contenu varié, ces dictionnaires

sont des fichiers créés et maintenus par des systèmes de gestion de bases de données. L'adoption d'une telle configuration permet une analyse plus précise, même si elle demande une travail antérieur majeur, comme celui de l'alimentation et de la mise à jour des dictionnaires.

Dans la deuxième étape, les textes sont soumis à un programme de parsage qui reconnaît les chaînes morphosyntaxiques propres aux UTC définies par la grammaire. L'interprétation de la représentation interne de la séquence réalisée lors du parsage utilise l'information de la base de connaissances des prédications pour décider si la chaîne recueillie constitue ou non une UTC. Cette base est structurée en fonction des fichiers liés par des nœuds de nature grammaticale, morphosyntaxique et sémantique. La procédure de consultation de la base de connaissances des prédications doit permettre à l'usager d'actualiser l'information qu'elle contient étant donné les changements éventuels de structures terminologiques au fur et à mesure que la terminologie d'un domaine évolue.

Si le programme confirme l'existence de la structure de l'UTC, il passe à la troisième étape, c'est-à-dire à l'insertion de l'UTC dans une liste. Cette liste, fournie par le système, doit être considérée comme une liste de termes potentiels qui pourront être conservés ou éliminés selon des critères fondés sur des fonctions sémantiques et établis par le terminologue. Cette analyse des fonctions sémantiques des composants de l'UTC constitue la quatrième étape de notre configuration. Finalement, la cinquième étape concerne l'extraction des UTC proprement dites accompagnées de leurs données morphosyntaxiques (p. ex. patrons morphosyntaxiques), de leurs fonctions sémantiques et de leurs relations conceptuelles.

La nature de notre recherche nous permet de fournir quelques pistes au niveau du contenu de la base de connaissances des prédications et de sa gestion par le logiciel. Ainsi, nous allons nous concentrer sur des spécifications d'ordre linguistique qui devront être prises en considération lors de la conception du système.

5.1.1.1 Pistes au niveau morphosyntaxique

Une fois que le système a étiqueté tous les mots du texte, le programme de parsage doit relever des séquences morphosyntaxiques selon des modèles terminologiques prédéfinis. Notre recherche a démontré l'existence des formules morphosyntaxiques suivantes :

- [[N [Adj.]] [Loc. Adv.]];
- [[N [Prép. + Dét. + N]] [Loc. Adv.]];
- [[N [Prép. + N]] [Loc. Adv.]];
- [[N [Prép. + N]] [Prép. + Dét. + N]];
- [N [Adj. [Prép. + N [Adj.]]]];
- [N [Adj. [Prép. + N [Prép. + N]]]];
- [N [Adj. [Prép. + N]]];
- [N [Adj.] [Adj. [Prép. + N]]].
- [N [Adj.] [Adj.] [Prép. + N]];
- [N [Adj.] [Adj.]];
- [N [Adj.] [Adv. [Adj.]]];
- [N [Adj.] [Prép. + N [Prép. + Dét. + N [Adj.];
- [N [Adj.] [Prép. + N]];
- [N [Adj.]];
- [N [Adv. [Adj.]]];
- [N [Loc. Adv.]];
- [N [Prép. + Dét. + N [Adj.]]];
- [N [Prép. + Dét. + N]];
- [N [Prép. + N [Adj.]] [Adj. [Prép. + N]]];
- [N [Prép. + N [Adj.]]];
- [N [Prép. + N [Adv. [Adj.]]]];
- [N [Prép. + N [Prép. + Dét. + N]]];

- [N [Prép. + N [Prép. + N [Prép. + N]]]];
- [N [Prép. + N [Prép. + N]]];
- [N [Prép. + N] [Adj.] [Adj.]];
- [N [Prép. + N] [Prép. + N]];
- [N [Prép. + N]];

Comme nous l'avons vu au chapitre 4, il s'agit de 27 modèles dont la fréquence varie dans le corpus. En effet, les modèles moins longs sont les plus souvent rencontrés. Nous remarquons également qu'il y a des matrices qui sont composées de la même séquence de catégories grammaticales, mais qui s'analysent de façon différente. Tel est le cas de :

- [N [Adj.] [Adj.] [Prép. + N]];
- [N [Adj.] [Adj. [Prép. + N]]].

Ce qui distingue ces deux formules morphosyntaxiques, ce sont les rapports sémantiques exprimés entre les composants. Ces rapports sont symbolisés par des crochets. Dans le premier cas, il y a trois arguments qui ont pour fonction de préciser le concept de la base de l'UTC. Chaque argument contient une ([Adj.]) ou deux catégories grammaticales ([Prép. + N]]) encadrées par des crochets. Dans le second cas, il y a deux arguments. Le premier est composé d'un élément de la catégorie grammaticale adjectivale ([Adj.]). Le second argument, plus complexe dans sa structure morphologique, est constitué de la séquence [Adj. [Prép. + N]]]. Ce dernier représente les arguments qui contiennent une expansion.

Dans le cas de la formule [N [Adj.] [Adj. [Prép. + N]]], l'argument [Adj. [Prép. + N] caractérise un type d'expression figée du domaine de la biotechnologie des végétaux. Ces expressions devront faire partie d'une liste à part dans le logiciel puisque, une fois vérifiées dans le texte, il y a une forte probabilité qu'une UTC existe. Les arguments *livre de* (exempt de), *isento de* (exempt de) sont des exemples d'expressions

figées. Si ces arguments font partie d'une liste gérée par le système, le logiciel sera capable de trouver des UTC comme : *clone livre de vírus* (clone exempt de virus), *planta livre de doença* (plant exempt de maladie) ou *planta livre de vírus* (plant exempt de virus).

Les expressions latines telles que *in vivo*, *in vitro*, *in situ* sont un autre élément important pour le dépistage d'UTC. Tout comme les arguments *livre de* (exempt de), *isento de* (exempt de), celles-ci devront faire partie d'une liste à part. Ces locutions adverbiales ont été vérifiées dans notre corpus par les modèles

- [[N [Adj.]] [Loc. Adv.]];
- [[N [Prép. + Dét. + N]] [Loc. Adv.]];
- [[N [Prép. + N]] [Loc. Adv.]];
- [N [Loc. Adv.]].

La vérification de la séquence *préposition* plus *déterminant* dans les six modèles cités ci-dessous, nous amène à rejeter la possibilité d'insérer ce type de composant dans des anti-dictionnaires[59], à moins d'avoir un moyen de vérifier sa position à l'intérieur de la phrase, étant donné qu'une UTC ne peut ni débuter ni finir par ce composant.

- [[N [Prép. + Dét. + N]] [Loc. Adv.]];
- [[N [Prép. + N]] [Prép. + Dét. + N]];
- [N [Adj.] [Prép. + N [Prép. + Dét. + N [Adj.];
- [N [Prép. + Dét. + N [Adj.]]];
- [N [Prép. + Dét. + N]];
- [N [Prép. + N [Prép. + Dét. + N]]].

La catégorie adverbe a été recensée dans trois modèles morphosyntaxique, à savoir

[59] Recueil d'unités qui ne sont pas censées faire partie de la structure d'une UTC.

- [N [Adj.] [Adv. [Adj.]]];
- [N [Adv. [Adj.]]];
- [N [Prép. + N [Adv. [Adj.]]]].

Pour que le système retienne l'adverbe comme étant une composante de l'UTC, il faut que ce dernier soit suivi d'un adjectif et non d'un verbe. D'après les matrices morphologiques citées ci-dessus, l'ensemble *adverbe* plus *adjectif* suit une base-nom, un argument-adjectif ou un argument de type syntagme prépositionnel. Le fait que ces adverbes finissent toujours par le suffixe *-mente* (-ment), peut constituer un bon indice pour dépister automatiquement ce type de séquence.

L'extraction fondée sur des modèles morphosyntaxiques est intéressante puisqu'elle emploie des techniques qui effectuent la recherche non pas par une séquence des mots, mais sur une séquence de codes qui représentent les catégories grammaticales de l'UTC. Toutefois, dans certains cas la représentation par mot peut augmenter les chances de trouver de nouvelles UTC. Étant donné que plusieurs systèmes précisent les prépositions qui peuvent figurer dans les structures morphologiques de type [N [Prép. + N]], nous croyons qu'il serait intéressant d'appliquer cette méthode à la structure de type [N [Adj.]]. Ainsi, nous spécifierons les formants qui entrent dans la composition des UTC qui relèvent de ce modèle. Nous avons d'ailleurs constaté que dans notre corpus le formant *não* entre souvent dans la formation de l'argument de la matrice [N [Adj.]]. Dans l'axe syntagmatique de l'UTC, ce préfixe de négation est toujours inséré entre le nom et l'adjectif, contrairement à ce qui se produit au niveau de la phrase où le *não* est toujours placé devant un verbe et ainsi considéré comme un adverbe de négation. Ces différences constituent un indice à considérer dans la conception d'un programme d'extraction. De plus, il s'agit d'un paramètre facile à implémenter puisqu'il est fondé sur la vérification préalable d'une séquence de catégories grammaticales par le programme d'étiquetage. La présence d'un trait d'union dans certaines séquences terminologiques formées par le préfixe *não* constitue un autre point de repère de cette structure.

Vu que la structure [N [Adj.]] peut avoir une variante terminologique de type [N [Prép. + N]], et que ce phénomène est assez fréquent, nous proposons que le système cherche la forme correspondante de type [N [Prép. + N]] lorsqu'il identifie la première structure [N [Adj.]] ou l'inverse. Ainsi, quand le système trouve l'UTC *agregado celular* (agrégat cellulaire), il fait une recherche à partir de la même base nominale mais sur la matrice [N [Prép. + N]] pour trouver la variante *agregado de células* (agrégat de cellules).

Le fait que le système détecte la séquence *nom* + ***não*** + *adjectif* lui permet de repérer la forme [N [Adj.]] correspondante, ce qui constitue une piste supplémentaire pour l'élaboration d'un logiciel. De cette manière, si le système trouve l'UTC *tecido não meristemático* (tissu non méristématique) il y a une forte probabilité qu'il trouve aussi l'UTC *tecido meristemático* (tissu méristématique).

Toutes ces pistes constituent des points essentiels pour la conception d'un logiciel. Nous expliquerons plus en détail dans la section qui suit les critères sémantiques qui devront être pris en compte pour ajouter du poids aux pistes déjà explorées.

5.1.1.2 Pistes au niveau sémantique

Les pistes sémantiques mentionnées dans cette section sont fondées sur le fait que nous croyons que les informations sémantiques peuvent aider le système à évaluer la pertinence des séquences syntaxiques qu'il relève. Ainsi, ces pistes sont employées pour délimiter un sous-ensemble de termes potentiels. Ces pistes aident également le terminologue dans la mesure où elles proposent un processus qui, en bout de ligne, regroupe diverses informations utiles au traitement des données linguistiques pour formuler des définitions, discuter avec le spécialiste du domaine ou raffiner une recherche terminologique.

Une fois que le système a produit la liste de termes potentiels, il peut déterminer de façon semi-automatique, les fonctions sémantiques de chaque composant de l'UTC et, par conséquent, indiquer automatiquement les relations conceptuelles qui en résulte. Une analyse morphosémantique automatique de chaque composant de l'UTC lié à un choix humain aide davantage à déterminer les fonctions sémantiques et, ainsi, à formuler les relations conceptuelles entre les composants de l'UTC. Par exemple, nous avons montré tout au cours du chapitre précèdent que la relation conceptuelle *agit sur* exige une base *processus* (un processus qui agit sur) ou *agent* (un agent qui agit sur) et un argument *patient* (patient qui subit l'action de l'agent ou du processus). Ainsi, si le système trouve, à partir d'une analyse morphosémantique, une base qui prend le suffixe *-mento* (-ment) ou *-ção*, il sait qu'il s'agit d'un *processus* puisque ces suffixes sont utilisés pour former des mots qui expriment cette notion. La fonction sémantique de l'argument est indiquée par le terminologue à partir d'une fenêtre qui montrera seulement les fonctions sémantiques des arguments qui ont un rapport avec une base *processus* et qui se restreint à une règle de formation déterminée (dans ce cas, la règle [ϕ [x_1]]). Si l'argument est un *patient* le système nous fournira automatiquement la relation conceptuelle *agit sur*.

Nous avons analysé deux paragraphes supplémentaires d'un de nos textes (v. annexe) et nous avons testé nos pistes de façon virtuelle pour vérifier la fiabilité de notre proposition. Dans notre texte de référence, il y avait 52 segments qui se caractérisaient comme des UTC potentielles selon les patrons morphosyntaxiques. En adoptant notre méthode de raffinement fondée sur le rôle sémantique de chaque composant de l'UTC, le système a accepté 17 UTC et a en rejeté 26. Les autres segments constituaient des silences (4), des bruits (3) et des exceptions (2). Nous allons examiner en détail chacun de ces résultats.

Les séquences qui ont été acceptées sont celles qui ont reçu des fonctions sémantiques qui étaient déjà prévues par le système, c'est-à-dire qui faisaient partie de la base de connaissance de prédication (v. figure 5.1 *supra* p. 229). Cet ensemble de fonctions sémantiques et de relations conceptuelles correspondantes résulte des tableaux présentés à

la fin de l'analyse de chaque règle de formation au chapitre précèdent. Nous citons ci-dessous quelques UTC retenues par le système.

Tableau 5.1 — UTC acceptées par le système

UTC	**Séquence des fonctions sémantiques**	**Relation conceptuelle résultante**
melhoramento genético (2 fois) (amélioration génétique)	*processus* + *appartenance*	*appartient à*
clonagem de genes (clonage de gènes)	*processus* + *résultat*	*produit un résultat*
genes de interesse (gène d'intérêt)	*entité* + *propriété*	*est du type*
seqüência de bases (séquence de base)	*groupement* + *contenance*	*contient*
espécies cultivada (espèces cultivées)	*entité* + *propriété*	*est du type*
heranças monogênicas (2 fois) (héritages monogéniques)	*entité* + *propriété*	*est du type*
braços de cromossomos (bras de chromosome)	*localisation* + *entité*	*se situe dans*
biologista molecular (biologiste moléculaire)	*entité* + *appartenance*	*appartient à*
porte de plantas (port de plantes)	*entité* + *appartenance*	*appartient à*
produção de leite (production de lait)	*processus* + *résultat*	*produit un résultat*
locus poligênicos (locus poligéniques)	*entité* + *contenance*	*contient*
locos de caractéristicas quantitativas (locus de caractéristiques quantitatives)	*entité* + *contenance* + *propriété*	*contient*
variações ambientais (variations environnementales)	*processus* + *patient*	*agit sur*
métodos biométricos (méthodes biométriques)	*processus* + *appartenance*	*appartient à*
variâncias genéticas (variances génétiques)	*processus* + *appartenance*	*appartient à*

Il faut préciser que la fonction sémantique de la base est attribuée automatiquement s'il s'agit d'une base active formée par des suffixes qui ont des fonctions sémantiques spécifiques quand ils sont ajoutés à cette base. Si la base est statique ou s'il s'agit d'une

base active qui comporte un suffixe que le système ne reconnaît pas, l'attribution de la fonction sémantique sera faite par le terminologue. C'est le terminologue qui attribue les fonctions sémantiques aux arguments puisque c'est le rôle sémantique de l'argument par rapport à son entourage terminologique qui est analysé et non uniquement la sémantique de l'argument à elle seule. Cependant, nous croyons qu'une étude plus approfondie de l'utilisation de la morphosémantique des arguments par rapport à la morphosémantique de la base, peut favoriser une systématisation pour l'automatisation.

Il convient aussi de souligner qu'on analyse une seule fois les UTC. Le logiciel reproduira de lui-même les codes qui se rapportent aux autres occurrences de cette même UTC. Tel est le cas de *melhoramento genético* (amélioration génétique) et *heranças monogênicas* (héritage mongénétique) qui ont été trouvés deux fois dans le texte.

En ce qui concerne le processus de rejet des séquences, nous le considérons également comme un facteur positif du système dans la mesure où le rejet diminue le bruit. Dans le système ici proposé, le segment est rejeté si le programme ne trouve pas une relation conceptuelle qui correspond à une séquence de fonctions sémantiques donnée. Par exemple, le système a rejeté le segment *modo de ação* (mode d'action) puisqu'il n'existait pas de relation conceptuelle correspondante pour la séquence *mode + processus*. Le rejet peut également être attribuable à la fonction sémantique Ø (nulle) de la base. En effet, il arrive que la base n'ait aucune des fonctions sémantiques déterminées dans notre étude, soit parce qu'elle est très générale ou qu'elle n'appartient pas au domaine étudié. Une fois que le terminologue a classé la fonction de la base comme étant nulle, le logiciel n'analyse plus le segment en question. Il passe alors au mot suivant dans le texte. S'il s'agit d'un argument qui reçoit la fonction sémantique Ø, le système analysera seulement la partie du segment qui a reçu des fonctions sémantiques.

Prenons à titre d'exemple le segment *casos de heranças monogênicas conhecidas*. Le système l'a retenu parce qu'il appartient à l'une des 27 matrices qui figurent dans la base de connaissance. Par la suite, le terminologue détermine la fonction sémantique de *casos*

que le logiciel lui présente comme étant la base du segment. Étant donné que *casos* est un mot qui n'appartient pas au domaine analysé, il a reçu la fonction sémantique Ø. Par conséquent, le logiciel ne va plus analyser *casos de heranças*. Il passera immédiatement au segment *heranças monogênicas conhecidas*. Le terminologue analyse cette nouvelle séquence et attribue une fonction sémantique Ø à l'argument *conhecidas*. Le système gardera donc uniquement la séquence correcte *heranças monogênicas*.

Il faut mentionner que l'attribution de fonctions sémantiques par le terminologue doit prendre en considération non seulement les rapports sémantiques entre les composants, comme nous l'avons mentionné ci-dessus, mais aussi le poids sémantique de l'UTC dans le domaine analysé. C'est d'ailleurs ce dernier critère qui rend l'automatisation complète du système difficile.

Lorsque le système rejette un bon segment, il s'agit d'un silence. Les quatre silences causés par notre méthode sont essentiellement des séquences dont la fonction sémantique ne faisait pas partie de notre corpus d'analyse. Dans ce cas, il faut prévoir une procédure de réalimentation de la base de connaissance à chaque fois que ce phénomène se produit.

Les deux exceptions sont des segments retenus à tort par le système, mais qui pourraient être évités par une analyse plus approfondie de la segmentation syntaxique au niveau de la phrase. Nous avons reproduit ci-dessous les contextes où figurent les deux exceptions mises en gras.

« [...] geralmente é abordado em **termos não convencionais** de [...] »
« [...] no desconhecimento, **via de regra** generalizado, do mode de [...] »

Il s'agit de locutions figées propres au discours en portugais et qui appartiennent à la langue générale. Le premier segment ***termos não convencionais*** est erroné puisqu'en fait *termos* s'analyse avec *em* pour donner la notion de manière (en termes de) et non pas avec ce qui suit. Dans le second segment *via de r**egra*** est une expression de la langue générale

qui signifie généralement. Nous proposons pour des cas de cette nature, d'alimenter un anti-dictionnaire qui filtrerait ces expressions de façon à les rejeter avant même le processus de marquage du texte.

5.2 Conclusion

Au cours de ce chapitre, nous avons proposé différentes stratégies pour l'amélioration du processus d'automatisation de l'extraction des UTC. Ces stratégies touchent tant l'aspect informatique que les aspects de nature essentiellement linguistique.

En ce qui concerne l'informatique, nous avons présenté une proposition d'environnement informatique basé sur l'intelligence artificielle et orienté vers la modélisation de connaissances linguistiques. Cette proposition est schématisée sous la forme d'une structure générale du traitement automatique pour l'extraction des UTC. Elle contient des étapes qui vont du marquage des mots dans le texte jusqu'à l'extraction des UTC. Un processus de validation avant l'extraction est prévu. Le système consulte constamment la base de connaissance alimentée par les résultats obtenus dans notre recherche. Nous avons également suggéré l'utilisation du langage Prolog pour la création du logiciel étant donné l'adéquation de ce langage de programmation au type de données que le système doit manipuler.

Pour ce qui est des aspects linguistiques, nous avons suggéré des pistes morphosyntaxiques et sémantiques. Ces pistes découlent de l'étude approfondie décrite au chapitre 4. Les pistes morphosyntaxiques font référence aux matrices morphosyntaxiques. Il s'agit des 27 matrices relevées dans la recherche qui servent de base pour l'exécution du programme de parsage dans le texte. Des caractéristiques morphosyntaxiques ont été mises en évidence pour quelques matrices afin de fournir des paramètres plus raffinés pour le parsage et ainsi obtenir une meilleure liste de termes potentiels, c'est-à-dire où les bruits et les silences sont réduits au minimum.

À cette liste de termes potentiels est appliquée une méthode semi-automatique de validation des UTC, ce qui constitue les pistes de nature sémantique. Il s'agit de l'attribution de la fonction sémantique à chaque composant du segment considéré comme valable par le système et de la vérification dans la base de connaissances de l'existence ou de l'absence de la séquence sémantique finale. Si la séquence existe, le programme extrait l'UTC et fournit, dans une fiche terminologique, l'étiquette qui exprime la relation conceptuelle de l'UTC ainsi que toute autre information utile pour le terminologue (p. ex. les catégories grammaticales, le contexte, etc.).

En guise de conclusion, nous voulons mentionner que même si les stratégies proposées dans le présent chapitre mettent en évidence des pistes de caractère essentiellement linguistique, nous croyons que les analyses d'ordre statistique peuvent également être ajoutées dans la conception du logiciel d'extraction des UTC. Toutefois, le concepteur du logiciel doit accorder la priorité aux méthodes linguistiques s'il veut obtenir les meilleurs résultats.

6. Conclusion

La reconnaissance, par l'ordinateur, des UTC en contexte est une tâche primordiale pour l'extraction terminologique automatisée. Pour que l'ordinateur réussisse à produire une liste de termes représentative d'un domaine à la fin de son processus d'extraction, il faut qu'il maîtrise adéquatement certains mécanismes linguistiques propres aux UTC. Ces mécanismes de nature plutôt complexe pour l'analyse humaine, s'avèrent tout aussi cruciaux pour l'analyse automatique. Cette complexité résulte principalement de la caractéristique dénominative des termes. La terminologie, une science onomasiologique, nomme les objets selon leur rôle dans un monde donnée. Cette nominalisation ou dénomination est représentée par des termes qui véhiculent le concept de fonction. La dénomination terminologique est, par conséquent, strictement liée à la fonction d'un mot à l'intérieur d'une UTC. C'est peut-être cette caractéristique qui rend l'extraction automatisée si difficile.

C'est en tenant compte du caractère fonctionnel de la terminologie que nous avons choisi la théorie de la grammaire fonctionnelle comme méthode d'analyse pour la présente recherche. Le grand avantage de cette grammaire est de systématiser les trois types de fonctions qui concernent les UTC, à savoir : la fonction sémantique, la fonction syntaxique et la fonction pragmatique. Mais, c'est surtout la méthode de la prédication proposée par cette grammaire, et que nous avons adaptée pour cette recherche, qui a véritablement rendu possible l'interprétation cohérente de nos données. En effet, par le biais de la méthode de la prédication, nous avons analysé les UTC comme des segments qui contiennent une base, des arguments et des satellites. La base et les arguments forment la prédication nucléaire.

La prédication nucléaire et les satellites constituent la prédication étendue. Cette façon de visualiser le segment terminologique facilite l'analyse des rapports linguistiques entre les composants des UTC tout en tenant compte de leurs fonctions à l'intérieur d'un domaine de connaissance.

À partir de cette méthode, nous avons défini des règles de formation des UTC. L'application de ces règles à nos données a permis de vérifier, à un niveau plus abstrait, dix modes de formation des UTC de notre corpus, soit

- [ϕ [x_1]];
- [[ϕ [x_1]] [y_1]];
- [ϕ [y_1]];
- [ϕ [x_1 [$x_{1/1}$]]];
- [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]];
- [ϕ [x_1] [x_2]];
- [ϕ [x_1] [x_2 [$x_{1/2}$]]];
- [ϕ [x_1] [x_2 [$x_{1/2}$ [$x_{1/1/2}$]]]];
- [ϕ [x_1 [$x_{1/1}$]] [x_2 [$x_{1/2}$]]];
- [ϕ [x_1] [x_2] [x_3]].

Cette typologie de la formation terminologique est représentée par un nombre variable d'UTC. Le mode [ϕ [x_1]], par exemple, représente 47,61 % de la totalité de nos donnée, c'est-à-dire presque la moitié du corpus. Étant donné que notre recherche traite exclusivement des UTC du domaine de la biotechnologie des végétaux, il serait très intéressant de reprendre cette typologie et de la tester sur d'autres domaines pour vérifier :

a) la présence ou l'absence de ces dix modes de formation;
b) la représentativité, en termes de fréquence, de chacun des modes de formation;
c) l'existence éventuelle d'autres modes de formation.

Les dix règles listées ci-dessus ont été interprétées sur les plans morphologique, syntaxique, sémantique et pragmatique, ce qui a fait ressortir des caractéristiques importantes pour la conception d'un logiciel d'extraction des UTC.

C'est à partir de ces caractéristiques que nous avons pu proposer de nouveaux éléments pour l'extraction terminologique. Nous avons ainsi suggéré un processus d'extraction semi-automatisé fondé sur la connaissance morphosyntaxique et sémantique des UTC. Ce processus fonctionne dans une structure de traitement terminologique qui privilégie le recours à l'intelligence artificielle. Dans la structure du traitement, les étapes suivantes sont prévues :

a) préparation des textes pour l'automatisation;
b) traitement du contenu par un marqueur de catégories grammaticales et un lemmatisateur;
c) analyse des textes par un programme de parsage;
d) présentation d'une liste de termes candidats;
e) validation de la liste de termes candidats.

Dans ce contexte, nous avons fourni d'importantes pistes pour améliorer les résultats produits par le logiciel d'extraction. Ces pistes servent à aider l'analyse syntaxique effectuée par le programme de parsage et à appuyer, par des moyens sémantiques, la validation de la liste de termes candidats. Quant au degré d'automatisation de l'implantation de ces pistes, les stratégies au niveau morphosyntaxique sont automatiques tandis que les stratégies au niveau sémantique ont été établies, pour le moment, d'une façon semi-automatisée. Nous envisageons sérieusement de réaliser une étude supplémentaire pour automatiser entièrement cette étape.

Parmi les caractéristiques de nature linguistique analysées dans notre recherche, nous voulons souligner celles que nous considérons essentielles pour l'automatisation tout en indiquant quelques perspectives pour des études futures.

Comme nous l'avons vu au cours du chapitre cinq, la présence d'UTC formées de bases actives facilite la recherche de la fonction sémantique effectuée par le logiciel. À partir d'une analyse morphosémantique du suffixe ajouté au verbe sous-jacent à la base active, il est possible d'arriver à de bons résultats pour ce qui est de la fonction sémantique de la base de l'UTC. Ce processus constitue d'ailleurs l'une des nos pistes au niveau sémantique pour le raffinement de la liste de termes candidats fournie par le système. Toutefois, les UTC sont aussi formées de bases statiques. Ce type de base représente la grande majorité des UTC recueillies dans notre recherche. Comme ces bases statiques, dans la plupart des cas composées de termes du domaine, ne dérivent pas de verbes, elles ne peuvent pas reprendre le processus que nous avons élaboré pour les bases actives. Par conséquent, dans notre système, l'analyse de ces bases exige une intervention humaine. Il serait ainsi profitable que d'autres études se penchent sur le sujet de la fonction sémantique des bases statiques afin de trouver une façon de systématiser ce type de composants de l'UTC pour qu'il puisse être automatiquement pris en considération par un logiciel d'extraction.

Il est important de rappeler que les UTC analysées dans la thèse possèdent une quantité définie de fonctions sémantiques. L'information sur la combinaison de ces fonctions sémantiques et les relations conceptuelles correspondantes est systématisée tout au long de l'analyse des données. Cette information est utilisée comme piste pour épurer la liste de termes candidats fournie par le processus d'extraction automatisée. Nous souhaitons que ces fonctions sémantiques et les relations conceptuelles qui en dérivent soient prises en compte dans des études futures afin d'enrichir la typologie sémantique présentée dans notre recherche.

Nous croyons que d'autres études devraient approfondir les aspects relatifs au comportement morphosémantique de chaque composant de l'UTC afin de trouver comment s'organise l'emploi des suffixes à l'intérieur de l'UTC. Les résultats de cette étude permettront d'améliorer la performance du logiciel en ce qui concerne l'attribution automatique des fonctions sémantiques. De cette façon, le terminologue disposera d'une liste de termes candidats plus complète.

Nos résultats ont démontré que bien souvent les déterminants et les adverbes ne sont pas pris en compte dans les patrons de fouille des logiciels. Pourtant, il s'agit de deux aspects morphosyntaxiques qui ont leur importance et qui devront faire partie des recherches futures si on veut diminuer le silence produit par le système. D'une part, les déterminants, même s'ils ne forment pas la majorité de notre corpus, existent et ne devront pas être mis de côté. D'autre part, l'adverbe entre dans la structure de plusieurs UTC. Nous avons trouvé des adverbes comme *fotossinteticamente* (photosynthètiquement), *genotipicamente* (génotypiquement), *geneticamente* (génétiquement) et *sexualmente* (sexuellement) dans diverses UTC de notre corpus. Ces adverbes ont une importante charge sémantique à l'intérieur des UTC, ce qui pourrait constituer un aspect d'étude important pour la constitution de nouvelles pistes sémantiques utilisées par le logiciel d'extraction.

Il sera très utile pour l'avancement des pistes sur l'automatisation d'élargir l'étude sur les rapports sémantico-syntaxiques démontrés dans notre recherche. Dans ce contexte, les rapports entretenus par les satellites et la prédication nucléaire ainsi que les relations maintenues entre un argument et son expansion constituent encore des questions qui devront être examinées dans d'autres domaines. Dans notre étude, nous avons avancé principalement l'analyse et l'interprétation linguistique de cette question, toutefois ce phénomène n'a pu être approfondi dans la partie traitant de l'automatisation en raison du travail que cette étape représentait. Les recherches futures pourront, par exemple, se pencher sur l'aspect des accords grammaticaux entre les composants de l'UTC, ce qui pourrait être d'une grande utilité pour la reconnaissance automatisée.

Finalement, il faut mentionner l'importance de tester nos résultats sur d'autres langues, d'autres domaines, d'autres structures (comme des UTC à base adjectivale ou qui comportent des expansions verbales) et d'employer des textes provenants de divers niveaux de langue comme corpus de données.

Cette thèse s'insère dans le *Projet intégré pour l'implantation et la diffusion de la terminologie scientifique et technique au Brésil* (v. Faulstich et Café, 1994), ce qui constitue un avantage puisque nous aurons la chance de tester nos résultats et de faire les ajustements nécessaires pour répondre à nos besoins grâce aux prototypes qui seront conçus par l'équipe d'informaticiens de l'Instituto Brasileiro de Informação e Tecnologia (IBICT). Nous prévoyons travailler de façon conjointe avec l'équipe du Centro Lexterm de l'Université de Brasília, dirigée par la professeure Enilde Faulstich.

Même si notre souci majeur était de faire avancer les connaissances en terminologie, nous n'avons jamais eu la prétention de résoudre toute la problématique d'une analyse linguistique des UTC et de son automatisation. Toutefois, nous pensons avoir fourni des éléments novateurs en ce qui concerne la théorie, la méthodologie et l'informatisation de cette question.

En effet, le recours à la grammaire fonctionnelle comme base scientifique de cette recherche et à la méthode de la prédication de laquelle nous avons dérivé les règles de formation des UTC constituent un apport considérable pour la terminologie. L'application de cette méthode nous a permis d'établir des matrices morphosyntaxiques enrichies d'information sur les fonctions syntaxiques, sémantiques et pragmatiques. Cette façon de procéder constitue un avancement par rapport aux études qui présentent simplement une liste de séquences de catégories grammaticales employées par les UTC. Finalement, nous croyons qu'il sera important d'établir une grille d'analyse sémantico-syntaxique que le terminologue utilisera pour valider la liste de termes candidats fournie par le système.

Bibliographie[60]

Ouvrages cités

ABBOU, André et *al.*, 1987, *Les industries de la langue, Les applications industrielles du traitement de la langue par les machines*, Paris, Éditions Daicadif, vol. 1, 399 p.

ABEILLÉ, Anne et Philippe BLACHE, 1997, L'état de l'art : la syntaxe, *T.A.L. Traitement automatique des langues*, vol. 38, n° 2, p. 69 - 90.

ABRAHAM, Maryvonne, 1994, Formalisation de la sémantique lexicale des verbes circuler et éviter, *T.A.L. Traitement automatique des langues*, vol. 35, n° 1, p.131 - 145.

AUGER, Pierre, 1978, La syntagmatique terminologique, typologie des syntagmes et limites des modèles en structures complexes, Dans : *Table ronde sur les problèmes de découpage du terme, Congrès de l'Association internationale de linguistique appliquée*, 20 - 26 août 1978, Montréal, OLF, p. 11 - 26.

----------, 1989a, La terminotique et les industries de la langue, *Meta*, vol. 3, n° 3, p. 450 - 456.

----------, 1989b, Informatique et terminologie : revue des technologies nouvelles, *Meta*, vol. 34, n° 3, p.485 - 491.

----------, 1990, Terminographie et lexicographie assistées par ordinateur : état de la situation et prospectives, Dans : *Actes du colloque Les industries de la langue, Perspectives des années 1990*, 21 - 24 novembre 1990, Montréal, OLF, Société des traducteurs du Québec, tome II, p. 659 - 680.

[60] Les sigles, qui figurent entre crochets à la suite de certaines références, sont ceux utilisés dans le champ observations de la base de données.

----------, 1994, Les outils de la terminotique : typologie des logiciels d'aide à la terminologie et/ou d'automatisation de la chaîne de travail en terminographie, *Terminologie nouvelles*, n° 11, p. 46 - 52.

----------, 1998, Des IDL aux NITC ou le règne numérique privé, *Terminogramme*, n°. 84 - 85, p. 4 - 5.

AUGER, Pierre, Patrick DROUIN et Alain AUGER, 1996, Filtact : un automate d'extraction des termes complexes, *Terminologie nouvelles*, n° 15, p. 48 - 51.

AUGER, Pierre, Patrick DROUIN et Marie-Claude L'HOMME, 1991, Automatisation des procédures de travail en terminographie, *Meta*, vol. 36, n° 1, p. 121 - 127.

BADIA, Toni, 1996, Bancos de conocimientos, *Terminômetro*, numéro spécial : La terminología en España, p. 71 - 74.

BENVENISTE, Émile, 1974, *Problèmes de linguistique générale*, Paris, Gallimard, tome 2, 286 p. (Collection Tel).

BOULANGER, Jean-Claude, 1978, Commentaire de Jean-Claude Boulanger au rapport de Roger Goffin sur le découpage du terme à des fins lexicographiques : critères formels, sémantiques, quantitatifs et taxinomiques, Dans : *Table ronde sur les problèmes de découpage du terme, Congrès de l'Association internationale de linguistique appliquée*, 20 - 26 août 1978, Montréal, OLF, p. 169 - 182.

----------, 1985, *Initiation à la terminologie [Terminologie et terminographie]* (Note du cours TRD-14436, 1re partie), Université Laval.

----------, 1989, La place du syntagme dans le dictionnaire de langue, *Meta*, vol. 34, n° 3, p. 516 - 528.

----------, 1995, Les éléments de formation technolectaux dans les dictionnaires généraux monolingues, *Présence francophone*, n° 4, p. 81 - 111.

BOURIGAULT, Didier, 1993, Analyse syntaxique locale pour le réperage de termes complexes dans un texte, *T.A.L. Traitement automatique des langues*, vol. 34, n° 2, p. 105 - 117.

BOUTIN-QUESNEL, Rachel et *al*., 1985, *Vocabulaire systématique de la terminologie*. Québec, Publications du Québec, 38p. (Cahiers de l'Office de la langue française).

BRITO, Célia, 1994, Uma proposta funcionalista, *Abralin. Boletin da Associação Brasileira de Linguística*, vol. 15, p.80 - 87.

CABRÉ, Maria Teresa, 1993, *La terminología; teoría, metodología, aplicaciones*, Trad. castellana de Carlos Tebé, Barcelona, Editorial Antárdida / Empúries, 529 p.

CASTILHO, Ataliba Teixeira de, 1994, Um ponto de vista funcional sobre a predicação, *Alfa*, vol. 38, p. 75 - 95.

CERVONI, Jean, 1991, *La préposition; étude sémantique et pragmatique*, Paris, Duculot, 309 p. (Champs linguistique).

CONDAMINES, Anne, 1994, Terminologie et représentation des connaissances, *Didaskalia*, n° 5, p. 35 - 51.

CORBIN, Danielle et Martine TEMPLE, 1994, Le monde des mots et des sens construits, *Cahiers de lexicologie*, vol. 65, n° 2, p. 5-28.

CUNHA, Celso et Lindley CINTRA, 1985, *Nova gramática do português contemporâneo*, 2e. édition. Rio de Janeiro, Nova Fronteira, 724 p. [CUNHA]

DAILLE, Béatrice, 1993, Extraction automatique de terminologie monolingue, Dans : *Actes du colloque informatique et langue naturelle*, Nantes, p. 1 - 21.

----------, 1994a, *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*, Paris, Université Paris 7, 228 p. (Thèse de doctorat).

----------, 1994b, *Extraction de noms composés terminologiques du domaine des télécommunications*, Paris, Université Paris 7, 13 p.

----------, 1994c, *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*, Paris, Université Paris 7, 9 p.

----------, 1995, *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistics Filtering*, Paris, Université Paris 7., 47 p.

DAILLE, Béatrice, Benoît HABERT, Christian JACQUEMIN et Jean ROYAUTÉ, 1996, Empirical Observation of Term Variation and Principles for their Description, *Terminology*, vol. 3, n° 2, p. 197 - 257.

DANEŠ, František, 1987, On Prague School Functionalism in Linguistics, Dans : *Functionalism in linguistics*, Amsterdam / Philadelphia, John Benjamins, p. 2 - 38 (Linguistic & Literary Studies in Eastern Europe, 20).

DAVID, Sophie, 1993, Remarques à propos du mode de construction des unités de forme NN, *T.A.L. Traitement automatique des langues*, vol. 34, n° 2, p. 59 - 74.

DAVID, Sophie et Pierre PLANTE, 1990, Le progiciel Termino : de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes, Dans : *Actes du colloque Les industries de la langue, Perspectives des années 1990*, 21 - 24 novembre 1990, Montréal, OLF, Société des traducteurs du Québec, tome I, p. 71 - 88.

DAVIDSE, Kristin, 1987, M.A.K. Halliday's Functional Grammar and the Prague School, Dans : *Functionalism in Linguistics*, Amsterdam / Philadelphia, John Benjamins, p. 39 - 79. (Linguistics & Literary Studies in Eastern Europe, 20).

DE SCHAETZEN, Caroline, 1994, Un peu d'histoire, *Terminologies nouvelles*, n° 11, p. 60 - 62.

----------, 1997, Typologie des outils de terminotique, Dans : *Mélanges de linguistique offerts à Rostislav Kocourek*, s.l., Presses Alfa, p. 81 - 87.

DIK, Simon, 1978, *Functional Grammar*, North-Holland, 230 p. (North-Holland Linguistics Series, 37).

----------, 1980, *Studies in Functional Grammar*, London, Academic Press, 245 p.

---------- (éd.), 1983, *Advances in Funcional Grammar*, Foris Publications, 415 p. (Publications in Languages Sciences, 11).

----------, 1981, Predication and Expression: the Problem and the Theorical Framework, Dans : *Predication and Expression in Functional Grammar*, London, Academic Press, p. 1 - 17.

----------, 1987, Some Principles of Functional Grammar, Dans : *Funcionalism in linguistics*, Amsterdam / Philadelphia, John Benjamins, p.81 - 100. (Linguistics & Literary Studies in Eastern Europe, 20).

----------, 1989, *The Theory of Functional Grammar*, part I: The Structure of the Clause, Dordrecht - Holland / Provindence RI - USA, Foris Publications, 433 p. (Functional Grammar, 9).

----------, 1989a, FGCMNLU: Functional Grammar Computational Model of the Natural Language User, Dans : *Functional Grammar and the Computer*, Dordrecht - Holland / Provindence RI - USA, Foris Publications, 322 p. (Functional Grammar, 10).

DIRVEN, René et Vilém FRIED, 1987, By way of introduction, Dans : *Functionalism in linguistics*, Amsterdam / Philadelphia, John Benjamins, p. ix - xvii. (Linguistic & Literary Studies in Eastern Europe, 20).

DOWNING, Pamela, 1977, On the Creation and Use of English Compound Nouns, *Language*, vol. 53, n° 4, p. 810 - 842.

DROUIN, Patrick, 1997, Une méthodologie d'identification automatique des syntagmes terminographiques : l'apport de la description du non-terme, *Meta*, vol. 42, n° 1, p. 45 - 54.

DUBOIS, Jean et *al*., 1991, *Dictionnaire de linguistique*, Paris, Larousse.

DUCROT, Oswald et Tzvetan TODOROV, 1972, Fonctionnalisme, Dans : *Dictionnaire encyclopédique des sciences du langage*, Paris, Seuil, p. 42 - 48.

FAULSTICH, Enilde, 1995, *Base metodológia para pesquisa em socioterminologia;* termo e variação, Brasília (Brésil), UnB / LIV, 31 p.

----------, 1997, *Da lingüística histórica à terminologia*, Brasília (Brésil), UnB/Liv, 31 p.

----------, s.d.a, Considerações sobre a variação e sinonímia na constituição de terminologias, Brasília (Brésil), UnB/Liv, 6 p.

----------, s.d.b, Exemplos de variantes terminológicas no português do Brasil, Brasília (Brésil), UnB/Liv, 12 p.

FAULSTICH, Enilde et Ligia CAFÉ, 1994, *Projeto integrado para implantação e difusão de terminologia científica e técnica no Brasil*, Brasília (Brésil), UnB / LIV; IBICT / DTI, 115 p.

FEUILLET, Jack, 1991, Adjectifs et adverbes : essai de classification, Dans : *Les état de l'adverbe, Travaux linguistique du Cerlico*, p. 35 - 58.

GAUDIN, François, 1993, *Socioterminologie; des problèmes sémantiques aux pratiques institutionnelles*, Rouen, 255 p. (Publications de l'Université de Rouen, 182).

GEBRUERS, Rudi, 1987, S. C. Dik's Functional Grammar: a Pilgrimage to Prague?, Dans : *Functionalism in Linguistics*, Amsterdam / Philadelphia, John Benjamins, p. 101 -134. (Linguistic & Literary Studies in Eastern Europe, 20).

GOUADEC, Daniel, 1987, Les horizons de la terminotique, *Meta*, vol. 32, n° 2, p. 130 - 138.

GREVISSE, Maurice, 1986, *Le bon usage*, 12 éd., Paris, Duculot, 1768 p.

GROSS, Gaston, 1990, Les noms composés dans un dictionnaire électronique, Dans : *Actes du colloque Les industries de la langue, Perspectives des années 1990*, 21 - 24

novembre 1990, Montréal, OLF, Société des traducteurs du Québec, tome II, p. 681 - 702.

----------, 1996, *Les expressions figées en français, noms composés et autres locutions*, Paris, Ophrys, 161 p. (Collection l'essentiel français).

GUILBERT, Louis, 1965, *La formation du vocabulaire de l'aviation*, Paris, Librairie Larousse, 712 p. (Thèse de doctorat).

----------, 1970, La dérivation syntagmatique dans les vocabulaires scientifiques et techniques, Dans : *Les langues de spécialité, Actes du Stage de Saint - Cloud*, 23 - 30 novembre 1967, Strasbourg, Aidela, p. 116 - 126.

----------, 1975, *La créativité lexicale*, Paris, Librairie Larousse, 285 p. (Langue et langage).

GUIMIER, Gaston, 1996, *Les adverbes du français; le cas des adverbes en -ment*, Paris, Ophrys, 170 p. (Collection l'essentiel français).

HABERT, Benoît et Christian JACQUELIN, 1993, Noms composés, termes, dénominations complexes, *T.A.L. Traitement automatique des langues*, vol. 34, n° 2, p. 5 - 41.

HOEKSTRA, T. et *al*. (éd.), 1981, *Perspectives on Functional Grammar*, Dordrecht - Holland / Cinnaminson - USA, Foris Publications, 352 p.

HOFFMAN, Lothar, 1987, Syntactic Aspects of LSP, *Fachsprache*, vol. 4, n° 3, p. 98 - 105.

KAGEURA, Kyo et Bin UMINO, 1996, Methods of Automatic Term Recognition: a Review, *Terminology*, vol. 3, n° 2, p. 259 - 289.

KOCOUREK, Rostislav, 1979, Lexical Phrases in Terminology, Dans : *Travaux de terminologie*, Québec, Université Laval, n° 1, p. 121 - 153.

----------, 1991, *La langue française de la technique et de la science*; vers une linguistique de la langue savante, 2e éd., Wiesbaden, Oscar Brandstetter Verlag KG et Co, 327 p.

LADOUCEUR, Jacques, Guylaine COCHRANE, Ligia Maria CAFÉ DE MIRANDA, 1997, Une méthode pour l'alignement de termes complexes plurilingues dans des textes spécialisés, *TradTerm*, São Paulo (Brésil), Humanitas publicações, FFLCH-USP, v. 4, n° 2, p. 199 - 221.

LADOUCEUR, Jacques et Patrick DROUIN, 1997, Une analyse terminométrique pour le repérage automatique des descripteurs complexes dans les textes de spécialité, *Meta*, vol. 42, n° 1, p. 207 - 218.

LAURISTON, Andy, 1994, Automatic Recognition of Complex terms: Problems and the TERMINO Solution, *Terminology*, vol. 1, n° 1, p. 147 - 170.

LEONARD, Rosemary, 1984, *The Interpretation of English Noun Sequences on the Computer*, Amsterdam / Philadelphia, Elsevier Science, 435 p. (North - Holland Linguistic Series, 51).

LERAT, Pierre, 1994, Composé syntagmatique, dénomination, terminologie, *Cahier de lexicologie*, vol. 65, n° 2, p. 151 - 158.

LEVI, Judith N., 1978, *The Syntax and Semantics of Complex Nominals*, New York, Academic Press, 301 p.

L'HOMME, Marie - Claude, 1988, *Origine et développement des industries de la langue*, Québec, Ciral, 141 p. (Publication K-8).

----------, 1995, Processing Word Combinations in Existing Term Banks, *Terminology*, vol. 2, n° 1, p. 141 - 162.

----------, 1996, Sélection des prépositions dans les termes complexes N (prép.) N à partir de leur structure conceptuelle, *Cahier de lexicologie*, vol. 68, n° 1, p. 25 - 43.

----------, 1996a, Sélection des prépositions dans les termes complexes français en vue d'applications en traduction automatique, Dans : *Actes du Colloque Informatique & Langue Naturelle I.L.N.96*, 9 - 10 octobre 1996, Nantes, p. 383 - 393.

----------, 1996b, Analyse des termes complexes binominaux français et anglais : problèmes et éléments de solutions, *Alfa*, vol. 9, p. 81 - 92.

L'HOMME, Marie-Claude, Loubna BENALI, Claudine BERTRAND et Patricia LAUDUIQUE, 1996, Definition of an Evaluation Grid for Term-Extraction Software, *Terminology*, vol. 3, n° 2, p. 291 - 312.

L'HOMME, Marie-Claude et Jocelyn GAGNON, à paraître, Étude de quelques structures conceptuelles entrant dans la formation de termes complexes N (préposition) N, à paraître dans *Pesquisa Linguística.*

L'HOMME, Marie-Claude et René GEMME, 1997, Modèle d'accès informatisé aux combinaisons lexicales spécialisées : verbe + nom (terme) et extension au nom (déverbal) + préposition + nom (terme), Dans : *Mélanges de linguistique offerts à Rostislav Kocourek*, s. l., Presses d'Alfa, p. 89 - 103.

LYONS, John, 1981, *Language and Linguistics*, Cambridge, Cambridge University Press, 356 p.

MAHMOUDIAN, Mortéza, 1979, Présentation, Dans : *Linguistique fonctionnelle : débats et perspectives*, Paris, PUF, 312 p.

MARTINET, André, 1985, *Syntaxe générale*, Paris, Armand Colin, 266 p.

----------, 1991, *Éléments de linguistique générale*, Paris, Armand Colin, 221 p.

----------, 1994, Qu'est-ce que la linguistique fonctionnelle ?, *Alfa*, vol. 38, p. 11 - 18.

McENERY, Tony et Andrew WILSON, 1996, *Corpus Linguistics*, Edinburgh, Edinburgh University Press, 209 p. (Edinburgh Textbooks in Empirical Linguistics).

MEYER, Ingrid, Lynne BOWKER et Karen ECK, 1991, Construction a Knowledge-Based Term Bank : Fundamentals and Implications, Dans : *Actes du symposium international Terminologie et documentation dans la communication spécialisée (T&D 91)*, s.l., Infoterm, Secrétariat d'État du Canada, p. 232 - 256.

MIRANDA, Ligia Maria Café de et Guylaine COCHRANE, 1997, Unités terminologiques complexes et adverbes : des éléments compatibles ?, *Travaux du LILLA*, n° 2, p. 33 - 43.

MULLER, Charles, 1979, Quelques principes de linguistique statistique : expérience, calcul et prévision, Dans : *Langue française et linguistique quantitative. Recueil d'articles*, Genève, Editions Slatkine, p. 215 - 222.

NAULLEAU, Elie, Marie-Gaëlle MONTEIL et Benoît HARBERT, 1996, Process Terms, *Euralex'96, Proceedings I - II*, part. II, p. 807 - 816.

NEVES, Maria Helena de Moura, 1994, Uma visão geral da gramática funcional, *Alfa*, v. 38, p. 109 - 127.

----------, 1997, *A gramática funcional*, São Paulo, Martins Fontes, 160 p.

OTMAN, Gabriel, 1989, Terminologie et intelligence artificielle, *La banque des mots*, muméro spécial, p. 63 - 95.

----------, 1991, Des ambitions et des performances d'un système de dépouillement terminologique assisté par ordinateur, *La banque des mots*, n° 4, p. 59 - 96.

----------, 1997, Les bases de connaissances terminologiques : les banques de terminologie de seconde génération, *Meta*, vol. 42, n° 2, p. 244 - 256.

PERINI, Mário A., 1996, *Gramática descritiva do português*, 2 éd., São Paulo (Brésil), Ática, 380 p.

PERRON, Jean, 1989, Termino : un système de dépouillement terminologique, *Terminogramme*, n° 54, p. 3 - 9.

PUGH, Jeanette, 1984, Contrastive Conceptual Analysis of Noun Compound Terms in English, French and Spanish Within a Restricted, Specialised Domain, R.R.K. Hartmann (éd.), Dans : *Lexeter '83, Papers from the International Conference on Lexicography at Exeter*, 9 - 12 September 1983, Türbingen, Max Niemeyer Verlag, p. 395 - 400, (Lexicographica Maior, 1).

QUIRK, Randolph et *al*., 1972, *A Grammar of Contemporary English*, London, Longman, 1120 p.

RITZKE, Johannes, 1981, Problèmes de l'analyse automatique des prépositions du français, Dans : *Analyse des prépositions, IIIme Colloque franco-allemand de linguistique théorique*, 2 - 4 février 1981, Tübingen, Max Niemeyer Verlag, p. 139 - 157, (Linguistiche Arbeiten, 110).

RIVALLAND, Gilles, 1993, Application de la méthode des graphes conceptuels au traitement de textes scientifiques, *Travaux du cercle linguistique de Nice*, n° 15, p. 49 - 67.

ROUGES-MARTINEZ, Josiane et Jean-Louis FOSSAT, 1994, Implantation terminologique en télédétection aérospatiale : méthodologie et mesure, *Terminologies nouvelles*, n° 12, p. 123 - 135.

SAGER, Juan C., 1990, *A Practical Course in Terminology Processing*, Amsterdam / Philadelphia, John Benjamin Publishing Company, 254 p.

SAGER, Juan C., David DUNGWORTH et Peter F. MCDONALD, 1980, *English Special Languages: Principles and Practice in Science and Tecnhology*, Wiesbaden, Oscar Brandstetter Verlag, XXIII, 368 p.

SAGER, Juan. C et Kyo KAGEURA, 1994 / 1995, Concept Classes and Conceptual Structures: Their Role and Necessity in Terminology, Dans : *Actes de langue française et de linguistique, Terminologie et linguistique de spécialité; études de vocabulaires et textes spécialises, ALFA,* vols. 7 et 8, p. 191 - 215

SAMUELSDORFF, Paul O., 1989, Simulation of a Functional Grammar in Prolog, Dans : *FGCMNLU : Functional Grammar and the Computer*, Dordrecht - Holland / Provindence RI - USA, Foris Publications, p. 29 - 44. (Functional Grammar, 10).

SANDMAN, Antônio José, 1989, *Formação de palavras no português brasileiro contemporâneo*, Curitiba (Brésil), Scientia et Labor, Icône, 185p. [SANDMAN]

SOWA, John., 1984, *Conceptual Structures: Information Processing in Mind and Machine*, Massachusetts, Addison-Wesley, 481 p. (System Programming series).

WATTERS, John, 1985, The Place of Morphology in Functional Grammar: the Case of the Ejagham Verb System, Dans : *Predicates and Terms in Functional Grammar*, Dordrecht - Holland / Cinnaminson - USA, Foris Publications, p. 85 - 104. (Functional Grammar Series, 2).

ZINGLÉ, Henri, 1996, Outils et méthode d'extraction automatique de la terminologie, Dans : *Séminaire Realiter, Réflexions méthodologiques sur le travail en terminologie et en terminotique dans les langues latines*, Nice, p. 79 -91.

----------, 1997, Acquisition et traitement de données terminologiques avec ZTermino, *Travaux du Lilla*, n° 2, p. 7 - 22.

Ouvrages de référence

Par auteur

CAIGNON, Philippe et Amal JAMMAL, 1995, *Dictionnaire anglais-français et français-anglais des biotechnologies*, Québec, Edisen, Paris, Tec&Doc, 185 p. [DB]

COMMISSION OF THE EUROPEAN COMMUNITIES (CEC), 1990, *Biotechnology Glossary/ Glossaire de biotechnologie : English, français, Deutsch, italiano, nederlands, dansk, español, português, grec*, London, New York, El Servier, 1072 p. [BG]

EMBRAPA, 1996, *Base de dados. Pesquisa agropecuária. Acervo documental*.

FAULSTICH, Enilde, 1996, *Glossário de Biotecnologia - Genplant. Cultura de tecidos de plantas. Melhoramento genético de plantas*, Brasília (Brésil), Publicações de Brasília, 133 p. (Série Lexterm). [GBP]

LA MONTAGNE, Yves et Armand CORRIVEAU, 1982, *Glossaire des termes utilisés en génétique et amélioration des arbres forestiers*, Québec, Ministère de l'Énergie et des Ressources, 55 p. [GGA]

PARENT, Sylvain, 1990, *Dictionnaire des sciences de l'environnement*, Ottawa, Broquet, 748 p. [DES]

QUIRION, Pierre et Philippe BOURBEAU, s.d, *Lexique des sciences biologiques*, Québec, Coop Comptoir Sciences, 733 p. [LSB]

WALKER, John M., Michael COX et Allan WHITAKER, 1988, *The Language of Biotechnology; a Dictionnary of Terms*, Washington, ACS, 255p. [LB]

Par titre

LA BANQUE DE TERMINOLOGIE DU QUÉBEC (BTQ), 199?, Dans : *Le grand dictionnaire terminologique*, Québec, OLF. [BTQ]

NOVO DICIONÁRIO DA LÍNGUA PORTUGUESA, 1986, 2e édition, Rio de Janeiro (Brésil), Nova Fronteira, 1838 p. [AURÉLIO]

GLOSSÁRIO DE RECURSOS GENÉTICOS VEGETAIS, 1996, Brasília (Brésil), Embrapa, 62p. [GRGV]

LE NOUVEAU PETIT ROBERT, Dictionnaire alphabétique et analogique de la langue française, 1995, Paris, Dictionnaires Le Robert, 2552 p. [PETIT ROBERT]

TERMINOLOGIE DU GÉNIE GÉNÉTIQUE, ?1997, Paris, Centre de ressources informatiques de l'Université René Descartes. [TGG]

TERMIUM, banque du Bureau de la traduction, 1996, Canada, Gouvernement du Canada, Travaux publics et services gouvernementaux. [TERMIUM]

Ouvrages constituant le corpus

LUZ, J.M.Q., 1993, *Obtenção "in vitro" de plantas de mandioquinha salsa (Arracacia xanthorriza Bancroft) via cultura de meristemas*, Lavras (Brésil), ESAL, 52 p. (dissertation de maîtrise). [BRA]

MANTELL, S.H., J.A. MATTHEWS et R.A. McKEE, 1994, *Princípios de biotecnologia em plantas; uma introdução à Engenharia Genética de plantas*, trad. de João Lúcio de Azevedo, Margarida Lopes R. Aguiar-Perecin et Natal Antonio Vello, Ribeirão Preto (Brésil), Sociedade Brasileira de Genética, 333p. [PBP]

PINTO, A.C.Q., mar.1995, Melhoramento de mangueira (*Mangifera indica* L.) no ecosistema dos cerrados do Brasil central por meio de hibridação, *Pesquisa Agropecuária Brasileira*, Brasília (Brésil), vol. 30, n.3, p. 369 - 374. [BRA]

RESENDE, F. V., 1993, *Comportamento, em condições de campo, de plantas de alho (**Alho sativum** L.) obtidas por cultivo de meristemas*, Lavras (Brésil), ESAL, 63 p. (Dissertation de maîtrise). [BRA]

ROSA JUNIOR, E.J. et N.A.H. ZÁRATE, 1992, Efeito de 5 sistemas de preparo do solo sobre o desenvolvimento de 12 linhagens de pimentão (*Capsicum annuum* L.) e sobre características físicas e químicas de um latossolo roxo distrófico em Dourados, MS (Brésil), *Revista Científica e Cultural*, v. 7, n. 1, p. 31 -38. [BRA]

SOARES, A.A., M.A.P. RAMALHO et A.F. de SOUZA, 1994, Estimativa de progresso genético obtido pelo programa de melhoramento de arroz irrigado da EPAMIG na época de oitenta, *Pesquisa Agropecuária Brasileira*, Brasília (Brésil), v. 29, n.1, p. 97 - 104. [BRA]

TORRES, Antonio Carlos et Linda Styer CALDAS (éd.), 1990, *Técnicas e aplicações da cultura de tecidos de plantas*, Brasília (Brésil), ABCTP, Embrapa, 433p. [BRA1]

Texte utilisé pour attester les pistes de notre logiciel

MALUF, Wilson Roberto, Perspectivas da Aplicaçao da Biologia Molecular no Melhoramento de Plantas: O Uso dos RFLPs, 1990, Dans : *Torres, Antonio Carlos et Caldas, Linda Styer (éd.), Técnicas e aplicações da cultura de tecidos de plantas*, Brasília (Brésil), ABCTP, Embrapa, p. 381.

Annexes

Annexe 1 — Description des UTC

Liste des UTC ordonnée à partir du français

acclimatation des plantes au sol : aclimatação de plantas no solo
action de désagrégation : ação desagregadora
action endogène : ação endógena
action enzymatique : ação enzimática
activité biologique : atividade biológica
ADN hautement répétitif : DNA altamente repetitivo
ADN non répété : DNA não repetido
ADN non sélectionné : DNA não selecionável
agent biologique : agente biológico
agent de perméabilité : agente permebializante
agent de réduction : agente redutor
agent dénaturant : agente desnaturante
agent d'éparpillement et d'adhérence : agente espalhante adesivo
agent d'oxydation : agente oxidante
agent fusogène chimique : agente fusogênico químico
agent infectieux : agente infeccioso
agent infectieux : agente infectivo
agent intercalaire : agente intercalante
agent mutagène chimique : agente mutagênico químico
agent mutagène : agente mutagênico
agent osmotique : agente osmótico
agent pathogène : agente patogênico
agrégat cellulaire : agregado celular
agrégat de cellules : agregado de células
aisselle de la feuille : axila da folha
aisselle foliaire : axila foliar
altération chromosomique : alteração cromossômica
altération épigénétique : alteração epigenética
altération nucléaire : alteração nuclear
amplification génique : amplificação gênica
analyse génomique : análise genômica
anomalie génétique : anormalidade genética
anomalies structurelles : alterações estruturais
anthère in vivo : antera in vivo
anthère mature : antera madura
application biotechnologique : aplicação biotecnológica
arbre élite adulte : árvore elite adulta
ARNm hautement purifié : mRNA altamente purificado
arrangement chromosomique cryptique : arranjo cromossômico críptico
association symbiotique : associação simbiótica
autofécondation contrôlée : autofecundação controlada
auxine de synthèse : auxina sintética
bactérie à vie autonome fixatrice d'azote : bactéria de vida livre fixadora de nitrogênio
bactérie à vie autonome : bactéria de vida livre
bactérie anaérobie : bactéria anaeróbica

bactérie fixatrice d'azote : bactéria fixadora de nitrogênio
bactérie non virulente : bactéria não virulenta
base de feuille : base de folha
base de la feuille : base da folha
base génétique : base genética
biotechnologie végétale : biotecnologia de plantas
bourgeon apical : broto apical
bourgeon floral : botão floral
bulbe provenant de la multiplication conventionnelle : bulbo proveniente de multiplicação convencional
bulbille obtenue par multiplication conventionnelle : bulbilho obtido por multiplicação convencional
cal régénérateur stable : calo regenerativo estável
carte de restriction : mapa de restrição
carte de sites de restriction : mapa de sítios de restrição
carte génétique : mapa genético
cellule bactérienne hôte : célula bacteriana hospedeira
cellule en suspension : célula em suspensão
cellule germinative originelle : célula germinativa original
cellule hôte receveuse : célula hospedeira receptora
cellule hôte : célula hospedeira
cellule indifférenciée : célula não diferenciada
cellule interne hautement vacuolaire : célula interna altamente vacuolizada
cellule somatique génétiquement différente : célula somática geneticamente diferente
cellule somatique : célula somática
cellule végétale immobilisée : célula vegetal imobilizada
cellule végétale indéterminée : célula vegetal indeterminada
cellule végétale nourricière : célula vegetal nutriz
chaîne de résidus d'adénosine : cadeia de resíduos de adenosina
chloroplaste photosynthétiquement compétent : cloroplasto fotossinteticamente competente
chromosome non-apparié : cromossomo não-pareado
chromosome non-homologue : cromossomo não-homólogo
clonage de gènes de plantes : clonagem de genes de plantas
clonage rapide : clonagem rápida
clone exempt de virus : clone livre de vírus
clone sain exempt de virose : clone sadio livre de virose
colonie de cellules bactériennes : colônia de células bacterianas
composant génomique nucléaire de plantes : componente genômico nuclear de plantas
culture "in vitro" : cultivo "in vitro"
culture d'anthères in vitro : cultura de anteras in vitro
culture d'anthères : cultura de anteras
culture de tissus de plantes : cultura de tecidos de plantas
culture du méristème apical : cultura de meristema apical
culture en batch : cultura em batelada
culture en suspension : cultura em suspensão
culture in vitro : cultura in vitro
développement de la feuille : desenvolvimento foliar
dominance apicale : dominância apical
effet inhibiteur : efeito inibitório

élongation "in vitro" : alongamento "in vitro"
élongation cellulaire : alongamento celular
embryogenèse in vitro : embriogênese in vitro
embryogenèse somatique : embriogênese somática
embryoïde adventice bipolaire : embrióide adventício bipolar
embryon somatique mature : embrião somático maduro
embryon somatique : embrião somático
espèce hautement endogame : espécie altamente endógama
espèce sexuellement compatible : espécie sexualmente compatível
espèces sexuellement incompatibles : espécies sexualmente incompatíveis
expérimentation in vitro : experimento in vitro
extrémité de la tige : ápice do broto
famille de séquences hautement répétées : família de seqüências altamente repetitivas
famille de séquences modérément répétées : família de seqüências moderadamente repetitivas
gain génétique : ganho genético
gemme adventice : gemas adventícias
génome nucléaire : genoma nuclear
génome récepteur : genoma receptor
génotype de maturation précoce : genótipo de maturação precoce
génotype non inductif : genótipo não indutivo
génotype supérieur : genótipo superior
génotype transitoire : genótipo transitório
grains de pollen embryogéniques : grãos de pólen embriogênicos
hybridation in situ : hibridação in situ
hybridation somatique : hibridação somática
hybride nain : híbrido anão
hybride par complémentation : híbrido por complementação
hybride somatique : híbrido somático
individus génotypiquement identique : indivíduos genotipicamente idênticos
induction florale in vitro : indução floral in vitro
induction florale : indução floral
infection de cultivars par virus : infecção de cultivares por vírus
infection des plantes par virus : infecção de plantas por vírus
inhibiteur de croissance in vitro : inibidor de crescimento in vitro
initiation florale : iniciação floral
inoculation en milieu de culture : inoculação em meio de cultura
libération d'hybride pour la traduction : liberação de híbrido para tradução
lignée parentale supérieure : linhagem parental superior
marqueur génétique sélectable : marcador genético selecionável
marqueur génétique : marcador genético
matériel génétique exogène : material genético exógeno
matériel génétique nucléaire : material genético nuclear
mécanisme biologique de résistance : mecanismo biológico de resistência
mécanisme de l'action relative à l'hydrolyse : mecanismo da ação hidrolítica
mécanisme de régulation : mecanismo de regulação
mécanisme naturel d'incompatibilité : mecanismo natural de incompatibilidade
méristème apical : meristema apical
méthode de croisement multiple : método de cruzamento múltiplo
méthode de pollinisation mécanique : método de polinização mecânica

méthode de propagation artificielle : método de propagação artificial
méthode de propagation in vitro : método de propagação in vitro
méthode de propagation traditionnelle : método de propagação tradicional
méthode de propagation : método de propagação
méthode d'indexage de types d'hôtes : método indexador de tipos de hospedeiros
méthode enzymatique : método enzimático
méthode in vitro : método in vitro
méthode non enzymatique : método não enzimático
micro-organisme réducteur obligatoire : microrganismo redutor obrigatório
milieu de culture : ambiente de cultura
molécule d'ADN circulaire extrachromosomique : molécula de DNA circular extracromossômica
mutagenèse in vitro : mutagênese in vitro
mutagenèse in vivo : mutagênese in vivo
mutation génique simple : mutação gênica simples
mutation somatique : mutação somática
plant exempt de maladie : planta livre de doença
plant exempt de virus : planta livre de vírus
plante ?progénitrice féminine : planta progenitora feminina
plante ?progénitrice masculine : planta progenitora masculina
plante génotypiquement uniforme : planta genotipicamente uniforme
plante obtenue par multiplication conventionnelle : planta obtida por multiplicação convencional
plante originaire de la mulitiplication conventionnelle : planta oriunda de multiplicação convencional
plante originaire de la multiplication conventionnelle : planta originada de multiplicação convencional
plante provenant de la multiplication conventionnelle : planta proveniente de multiplicação convencional
pollinisation in vitro : polinização in vitro
processus de maturation naturelle : processo de amadurecimento natural
propagation clonale in vitro : propagação clonal in vitro
propagation clonale rapide : propagação clonal rápida
propagation clonale stable : propagação clonal com fidelidade ao tipo original
propagation clonale : propagação clonal
propagation en masse : propagação em massa
propagation in vitro : propagação in vitro
propagation végétative in vitro : propagação vegetativa in vitro
racines adventices : raízes adventícias
recombinaison homologue in vivo : recombinação homóloga in vivo
régénérant génétiquement modifié : regenerante geneticamente engenheirado
régénérant génétiquement stable : regenerante geneticamente estável
régénération in vitro : regeneração in vitro
réinfection artificielle : reinfecção artificial
réinfection naturelle par virus : reinfecção natural por vírus
réinfection virale : reinfecção viral
réinfection virale : reinfecção virótica
serre : casa de vegetação
site actif : sítio ativo
site de restriction : sítio de restrição
stabilité génétique : estabilidade genética

substance limitative de croissance : substância limitadora de crescimento
substance régulatrice de croissance : substância reguladora de crescimento
système de préparation de sol : sistema de preparo de solo
système de préparation du sol : sistema de preparo do solo
terre cultivée : área de solo cultivado
tissu différencié : tecido diferenciado
tissu embryonnaire : tecido embrionário
tissu hautement différencié : tecido altamente diferenciado
tissu méristématique : tecido meristemático
tissu multicellulaire : tecido multicelular
tissu non méristématique : tecido não meristemático
tissu somatique génétiquement distinct : tecido somático geneticamente distinto
tissu végétal mature : tecido vegetal maduro
traduction de l'ARN-m in vitro : tradução do mRNA in vitro
traduction réprimée par hybridation : tradução reprimida por hibridação
variabilité génétique naturelle : variabilidade genética natural
variabilité génétique : variabilidade genética
variété semi-naine : variedade semi-anã
vecteur de clonage de gènes de plantes : vetor de clonagem de genes de plantas
vecteur de clonage de gènes : vetor de clonagem de genes
vecteur de clonage : vetor de clonagem
vecteur de gènes de plantes : vetor de genes de plantas
vecteur génétique : vetor genético

Annexe 2 — Fonctions sémantiques

Fonctions sémantiques

action - fait de produire un effet ou d'agir sur une entité ou un processus.

Ex : *tradução* ***reprimida*** *por hibridação* (traduction réprimée par hybridation), *sítio de* ***restrição*** (site de restriction).

agent - entité qui produit une action.

Ex: ***agente*** *permebializante* (agent de perméabilité); ***substância*** *reguladora de crescimento* (substance régulatrice de croissance).

appartenance - situation qui lie une entité ou un processus à une collectivité, un ensemble ou un groupement.

Ex : *aplicação* ***biotecnológica*** (application biotechnologique), *tecido* ***vegetal*** *maduro* (tissu végétal mature).

but - fin pour laquelle une chose est faite ou existe.

Ex : *liberação de híbrido* ***para tradução*** (libération d'hybride pour la traduction).

composant - élément qui entre dans la composition d'une entité.

Ex : *planta* ***genotipicamente*** *uniforme* (plante génotypiquement uniforme).

contenance - situation qui représente le contenu d'une entité ou d'un groupement.

Ex : *agregado* ***de células*** (agrégat de cellules).

effet[61] - situation produite ou provoquée par un agent ou un processus.

Ex : *agente* ***infectivo*** (agent infectieux), *agente* ***espalhante adesivo*** (agent d'éparpillement et d'adhérence), *ação* ***desagregadora*** (action de désagrégation).

[61] L'effet est toujours un processus (ex : une infection, une adhérence, une désagrégation etc.)

entité - objet matériel ou abstrait.

Ex : ***antera*** *madura* (anthère mature), ***híbrido*** *anão* (hybride nain).

état - situation dans laquelle se trouve un objet abstrait ou matériel.

Ex : *antera* ***madura*** (anthère mature), *sítio* ***ativo*** (site actif).

forme - apparence extérieure qui donne à une entité sa spécificité, sa particularité.

Ex : *molécula de DNA* ***circular*** *extracromossômica* (molécule d'ADN circulaire extrachromosomique), *híbrido* ***anão*** (hybride nain).

groupement - association d'entités pour former un groupe, un ensemble ou une classe.

Ex : ***agregado*** *de células* (agrégat de cellules), ***colônia*** *de células bacterianas* (colonie de cellules bactériennes).

instrument - moyen par lequel s'accomplit un processus.

Ex : *agente fusogênico* ***químico*** (agent fusogène chimique), *agente mutagênico* ***químico*** (agent mutagène chimique), *célula somática* ***geneticamente*** *diferente* (cellule somatique génétiquement différente).

intensité - degré des propriétés d'une entité ou des façons de procéder.

Ex : *DNA* ***altamente*** *repetitivo* (ADN très répétitif), *mRNA* ***altamente*** *purificado* (ARNm hautement purifié), *tecido* ***altamente*** *diferenciado* (tissu hautement différencié), *propagação* ***em massa*** (propagation en masse).

localisation - endroit qui situe, circonscrit ou limite dans l'espace une entité ou un processus.

Ex : *aclimatação de plantas* ***no solo*** (acclimatation des plantes au sol), *cultura* ***in vitro*** (culture in vitro).

mode - manière d'être pour une entité ou d'effectuer un processus.

Ex : *processo de amadurecimento* ***natural*** (processus de maturation naturelle), *autofecundação* ***controlada*** (autofécondation contrôlée), *propagação clonal* ***rápida*** (propagation clonale rapide).

patient - entité qui subit une action ou un processus.

Ex : *iniciação* ***floral*** (initiation florale), *desenvolvimento* ***foliar*** (développement de la feuille), *vetor de clonagem de* ***genes*** (vecteur de clonage de gènes).

processus - événement qui modifie une situation initiale.

Ex : ***embriogênese*** *somática* (embryogenèse somatique), ***hibridação*** *somática* (hybridation somatique), ***propagação*** *vegetativa in vitro* (propagation végétative in vitro).

propriété - ensemble de constances, de caractères (surtout caractère de fonction), de réactions d'une substance; manière dont elle se comporte suivant les conditions dans lesquelles elle est placée.

Ex : *recombinação* ***homóloga*** *in vivo* (recombinaison homologue in vivo), *indivíduos genotipicamente* ***idênticos*** (individus génotypiquement identique), *linhagem* ***parental superior*** (lignée parentale supérieure).

résultat[62] - situation produite par un processus, un effet ou une action.

Ex : *indução* ***floral*** *in vitro* (indution florale in vitro), *propagação* ***clonal*** *in vitro* (propagation clonale in vitro).

source - origine d'une entité.

Ex : *híbrido* ***somático*** (hybride somatique)[63]

[62] Le résultat fait toujours référence à une entité (une fleur, un clone, etc.).
[63] L'hybride résulte des cellules somatiques.

Annexe 3 — Texte utilisé pour attester les pistes de notre logiciel

Texte utilisé pour attester les pistes de notre logiciel

Maluf, Wilson Roberto, Perspectivas da Aplicaçao da Biologia Molecular no Melhoramento de Plantas: O Uso dos RFLPs, 1990, Dans : *Torres, Antonio Carlos et Caldas, Linda Styer (éd.), Técnicas e aplicações da cultura de tecidos de plantas*, Brasília (Brésil), ABCTP, Embrapa, p. 381.

Introdução

O impacto da engenharia genética molecular no melhoramento genético geralmente é abordado em termos não convencionais de clonagem de genes desejáveis e sua inserção no genoma de um cultivar da mesma espécie ou de outra. Um pressuposto para o uso desta tecnologia é a possibilidade de identificar com precisão o gene ou genes de interesse e sua seqüência de bases. Essa identificação esbarra freqüentemente no desconhecimento, via de regra generalizado, do modo de ação desses genes, ou seja, no desconhecimento de seus produtos de transcrição e/ou tradução. Tomem-se como exemplos os inúmeros casos de genes para resistência a doenças em espécies cultivadas: em muitos casos, a genética e citogenética clássicas demonstram tratar-se de heranças monogênicas, que chegam a ser mapeadas com precisão a nivel de braços de cromossomos específicos, freqüentemente ligados por poucos centimorgans a marcadores fenotípicos de fácil avaliação. Toda essa massa de informações é freqüentemente de pouca valia para o biologista molecular na sua tarefa de clonar o gene de interesse.

É conveniente ressaltar, ainda, que os casos de heranças monogênicas conhecidas e mapeadas ao nível dos cromossomos constituem-se mais na exceção do que na regra do cotidiano do melhoramento genético animal ou vegetal: grande parte das características de interesse econômico (produtividade, resistência à seca, porte de plantas, produçao de leite, etc...) são características complexas, poligênicas e bioquimicamente indefinidas. São denominadas «características quantitativas» e os locus poligênicos envolvidos em sua expressão são os denominados «locos de características quantitativas» (do inglês «quantitative trait loci», ou, abreviadamente QTL). Esses caracteres são, em geral, sujeitos a variações ambientais. Sob o ponto de vista de melhoramento genético, estas características são abordadas, em geral, com métodos biométricos, que avaliam a magnitude das variâncias genéticas, ambientais e suas interações, e que podem tornar-se valiosas ferramentas nos programas de melhoramento genético.

Index des auteurs cités

Index des auteurs cités

Abbou, André et *al.*, 62

Abeillé, Anne et Philippe Blache 62

Abraham, Maryvonne, 49

Auger, Pierre, 45, 62, 63, 64, 90

Auger, Pierre, Patrick Drouin et Alain Auger 70, 78

Auger, Pierre, Patrick Drouin et Marie-Claude L'Homme 84

Badia, Toni, 86

Benveniste, Émile 44, 45, 56

Boulanger, Jean-Claude 42, 46, 53, 56, 59, 249

Bourigault, Didier 46, 76, 77

Brito, Célia 17

Cabré, Maria Teresa 63

Castilho, Ataliba Teixeira de 8, 28, 29

Cervoni, Jean 136

Condamines, Anne 87

Corbin, Danielle et Martine Temple 48

Cunha, Celso et Lindley Cintra 32

Daille, Béatrice 44, 74, 77, 121, 122

Daille, Béatrice, Benoît Habert, Christian Jacquemin et Jean Royauté 59

Daneš, František 6

David, Sophie 49, 53

David, Sophie et Pierre Plante 70

Davidse, Kristin 8

De Schaetzen, Caroline 66, 70, 81

Dik, Simon, 2, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 31, 37, 39, 132, 133, 170, 228, 252

Dirven, René et Vilém Fried 6

Drouin, Patrick 61, 71, 72, 75, 78

Dubois, Jean et *al.* 21, 32, 34

Ducrot, Oswald et Tzvetan Todorov 5

Faulstich, Enilde, 24, 29, 53, 60, 149, 217

Faulstich, Enilde et Ligia Café 89, 247

Feuillet, Jack 181, 182

Gaudin, François 98

Gebruers, Rudi 8

Gouadec, Daniel 83

Grevisse, Maurice 32, 34, 48

Gross, Gaston 43, 44, 47, 54, 56, 57, 58

Guilbert, Louis 45, 92, 95, 96, 125, 130, 144, 146, 147

Guimier, Gaston 34, 35, 36, 139, 170, 172, 178, 180, 219

Habert, Benoît et Christian Jacquelin 47, 48, 49, 55, 70, 71, 77, 78, 80

Hoekstra, T. et *al*. 9, 11

Hoffman, Lothar 122

Kageura, Kyo et Bin Umino 74

Kocourek, Rostislav 45, 46, 55, 56, 57, 58, 59, 97, 121, 122, 146, 176, 251, 254

L’Homme, Marie-Claude et Jocelyn Gagnon 52
L’Homme, Marie-Claude, Loubna Benali, Claudine Bertrand et Patricia Lauduique 68, 72, 74, 79
Ladouceur, Jacques et Patrick Drouin 68, 78
Lauriston, Andy 61, 65, 72
Leonard, Rosemary 51
Lerat, Pierre 55
Levi, Judith N. 51
Lyons, John 6
Mahmoudian, Mortéza 7
Martinet, André 7, 43
McEnery, Tony et Andrew Wilson 75
Meyer, Ingrid, Lynne Bowker et Karen Eck 85, 87
Miranda, Ligia Maria Café de et Guylaine Cochrane 178
Muller, Charles 76
Naulleau, Elie, Marie-Gaëlle Monteil et Benoît Harbert 83
Neves, Maria Helena de Moura 2, 11
Otman, Gabriel 61, 63, 67, 69, 72, 74, 82, 87, 88
Perini, Mário A. 162, 211
Perron, Jean 66, 68, 69, 70
Pugh, Jeanette 50
Ritzke, Johannes 138
Rivalland, Gilles 86
Sager, Juan C. 46, 64, 146, 147, 167
Sager, Juan C., David Dungworth et Peter F. McDonald 123
Sager, Juan. C et Kyo Kageura 50
Samuelsdorff, Paul O. 228
Sowa, John 30, 86
Zinglé, Henri 73, 81, 88, 228