Une approche par surclassement pour le contrôle d'un processus d'analyse linguistique

Grégory Smits

UNIVERSITÉ de CAEN/BASSE-NORMANDIE

U.F.R. : Sciences

ÉCOLE DOCTORALE : SIMEM

THÈSE

présentée par

Grégory SMITS

et soutenue

le 5 mai 2008

en vue de l'obtention du

DOCTORAT de l'UNIVERSITÉ de CAEN

spécialité : Informatique

(Arrêté du 7 août 2006)

Une approche par surclassement pour le contrôle d'un processus d'analyse linguistique

MEMBRES du JURY

Pierre Zweigenbaum	**Directeur de recherche**	**LIMSI-CNRS (rapporteur)**
Denis Bouyssou	**Professeur**	**Université de Paris Dauphine (rapporteur)**
Jacques Vergne	**Professeur**	**Université de Caen (directeur de thèse)**
Christine Chardenon	**Ingénieur de recherche**	**France Télécom R&D Lannion**
Gérard Sabah	**Directeur de recherche**	**LIMSI-CNRS, Paris**
Abdel Illah Mouaddib	**Professeur**	**Université de Caen**

Mis en page avec la classe thloria.

Remerciements

Je préfère amplement remercier les personnes concernées de vive voix.

À Zia : source de retard, d'inspiration et d'émerveillement ...

Table des matières

Table des figures xi

Liste des tableaux xiii

Introduction xv

Chapitre 1
Les processus de TALN et la génération des hypothèses concurrentes 1

1.1 Le traitement automatique de textes écrits 2
1.1.1 TALN : définition et enjeux 2
1.1.2 Caractériser le TALN par ses difficultés 3
1.1.3 Niveaux d'analyse 5
1.1.4 Les systèmes de TALN : entre inspirations psycholinguistiques et modèles informatiques 8
1.2 Le TALN dans un contexte industriel 10
1.2.1 Des techniques et outils d'analyse linguistique au service des applications 11
1.2.2 `TiLT` : une boîte à outils linguistique 12
1.2.3 Description détaillée d'un processus d'analyse réalisé par `TiLT` 14
1.3 Hypothèses concurrentes et ambiguïtés artificielles : un problème récurrent en TALN 17
1.3.1 Indéterminisme et génération d'erreurs d'interprétations 17
1.3.2 Quelques pistes d'explication de ce phénomène 19
1.3.3 Illustrations de la génération d'hypothèses concurrentes 20

Chapitre 2
Vers une approche décisionnelle de contrôle basée sur plusieurs critères de comparaison

2.1 Contrôler les hypothèses générées 26
2.1.1 Définition du contrôle : évaluer puis décider 26
2.1.2 Objectifs du contrôle 27

2.1.3 Des “points d’embarras” aux “points de décision” . . . 29
2.2 Évaluer la pertinence des hypothèses à partir d’informations distinctives complémentaires . . . 30
2.2.1 Différencier les hypothèses par observation de corpus . . . 30
2.2.2 Différencier les hypothèses à l’aide de sources de connaissances supplémentaires ou d’heuristiques . . . 31
2.2.3 Les "éléments de décision" dans TiLT . . . 32
2.2.4 Des "éléments de décision" aux critères de comparaison . . . 34
2.3 Une meilleure exploitation des critères de comparaison . . . 35
2.3.1 Limite des approches monocritère de contrôle . . . 36
2.3.2 Limites des approches multicritère de contrôle . . . 37
2.3.3 Décision ou aide à la décision : vers plus de généricité et de flexibilité . . . 40

Chapitre 3
L’aide multicritère à la décision et le TALN

3.1 L’Aide MultiCritère à la Décision (AMCD) . . . 44
3.1.1 L’AMCD en tant qu’extension pragmatique de la théorie de la décision . . . 44
3.1.2 Formalisation d’un problème d’AMCD . . . 45
3.1.3 Les approches en AMCD . . . 46
3.2 Les approches par surclassement . . . 49
3.2.1 Modélisation des préférences et connaissances expertes . . . 49
3.2.2 Le surclassement comme relation générique de comparaison . . . 51
3.2.3 ELECTRE III et ELECTRE TRI . . . 53
3.3 Une approche par surclassement pour le contrôle des indéterminations . . . 62
3.3.1 Particularités du contexte décisionnel . . . 62
3.3.2 Une méthodologique axée sur la généricité . . . 63
3.3.3 Adaptation et simplification d’aspects techniques . . . 63

Chapitre 4
Implémentation d’un système décisionnel de contrôle

4.1 Vers une architecture décisionnelle de contrôle . . . 69
4.1.1 Confronter les besoins de contrôle au processus d’analyse . . . 69
4.1.2 Architecture logicielle de TiLT : abstraire pour généraliser . . . 73
4.1.3 Modélisation du module décisionnel . . . 73
4.1.4 Externaliser pour adapter . . . 77
4.1.5 Proposition d’adaptation de l’architecture de traitement en vue d’une externalisation complète des aspects décisionnels . . . 78

4.2 Processus de contrôle . . . 80
4.2.1 Construction du contexte décisionnel . . . 80
4.2.2 Construction des relations de surclassement en vue d'un classement ou d'une sélection . . . 82
4.2.3 Construction des relations de surclassement en vue d'un tri . . . 85
4.2.4 Quelques fonctionnalités supplémentaires liées à la manipulation des critères 85
4.3 Suggérer un modèle de préférences à partir d'un corpus d'hypothèses de références 88
4.3.1 Regard statistique sur le modèle de préférences . . . 88
4.3.2 L'annotation comme l'expression des préférences d'un expert (décideur) . 89
4.3.3 Quelques heuristiques pour suggérer un modèle de préférences à partir d'un corpus de référence . . . 92

Chapitre 5
Évaluation de l'apport de l'AMCD pour le contrôle d'une chaîne de TALN

5.1 Classement des rubriques métier pour une requête soumise à un service d'annuaire de professionnels . . . 103
5.1.1 Le processus d'indexation de TiLT et l'apparition des indéterminations . . 103
5.1.2 Un "point de décision" déjà établi . . . 104
5.1.3 Vers une meilleure exploitation des critères de comparaison . . . 106
5.1.4 Résultats et interprétations . . . 108
5.2 Identification des cas de reprise anaphorique : problématique de tri . . . 110
5.2.1 Présentation du problème . . . 111
5.2.2 Présentation de la démarche expérimentale . . . 116
5.2.3 Interprétation des résultats : propriétés linguistiques des reprises anaphoriques . . . 127
5.3 Quelle transcription pour un SMS ? problématique de sélection . . . 130
5.3.1 Le processus symbolique de transcription des SMS . . . 130
5.3.2 Caractérisation et quantification des indéterminations . . . 133
5.3.3 Vers un contrôle statistique du processus d'analyse . . . 139
5.3.4 Application de la stratégie de contrôle, résultats et interprétations . . . 145

Conclusion **153**

Apports, perspectives et conclusion

1 Une intersection novatrice entre deux domaines de recherche . . . 156
1.1 Apport de notre approche pour la problématique étudiée . . . 156
1.2 Apport de l'AMCD pour le TALN . . . 158

1.3 Apport du TALN pour l'AMCD . . . 159
2 Perspectives . . . 160
2.1 Étendre le contrôle aux autres "points d'embarras" . . . 160
2.2 Compléter l'approche et poursuivre son évaluation . . . 162
2.3 Extensions et pistes de recherches envisageables . . . 163
3 Conclusion . . . 165
3.1 Légitimité des stratégies de contrôle . . . 165
3.2 Une approche de contrôle décisionnelle basée sur de multiples critères de comparaison . . . 166
3.3 Un système de contrôle opérationnel et fonctionnel . . . 167
3.4 Quelques expérimentations parmi l'ensemble des cas de contrôle envisageables . . . 169

Annexes **171**

Annexe A Rappels méthodologiques **171**
A.1 Formalisation commune des éléments décisionnels . . . 171
A.2 Formalisation des préférences . . . 171
A.2.1 Préférences orientées "outputs" . . . 171
A.2.2 Préférences orientées "inputs" . . . 172
A.3 Construction des relations de surclassement . . . 172
A.4 Compléments sur la méthode `ELECTRE III` : classement et sélection . . . 174
A.5 Compléments sur la méthode `ELECTRE TRI` : tri . . . 175

Annexe B `Beslissing` : un module de contrôle décisionnel **179**
B.1 Diagramme de classes détaillé du module de contrôle `Beslissing` . . . 179
B.2 Fichier de configuration commenté du module de contrôle `Beslissing` . . . 181

Annexe C Outils connexes au module décisionnel `Beslissing` **183**
C.1 Interface de configuration du module décisionnel `Beslissing` . . . 183
C.2 `CorpusTagger` : une interface d'annotation des hypothèses de référence et de construction des tables de performances . . . 185

Annexe D Expérimentation autour de la résolution de la coréférence **189**
D.1 Extraits du corpus "Le Monde" utilisé pour la résolution des liens de coréférence . 189
D.2 Critères d'évaluation des couples candidats antécédent/reprise . . . 190
D.3 Modèles de préférences . . . 192
D.3.1 Modèles de préférences définis *a priori* par un expert . . . 192
D.3.2 Modèles de préférences suggérés . . . 195

D.3.3 Modèles de préférences mixtes 199

Annexe E Expérimentation autour de la transcription de SMS **203**

E.1 Extraits du corpus utilisé pour la transcription de SMS 203

Annexe F Expérimentation autour de l'indexation de rubriques métiers **205**

F.1 Extraits du corpus de requêtes utilisées pour l'évaluation 205

F.2 Sortie `XML` de `TiLT` pour l'indexation de la requête "sécurité sociale" 206

Bibliographie **209**

Table des figures

1.1 Le TALN : un ensemble d'outils d'interactions et de communications "artificielles" 3
1.2 Représentation arborescente d'une analyse syntaxique en dépendance 6
1.3 Modularité des processus de TALN 9
1.4 `TiLT` : la boîte à outils linguistique développée par l'équipe `LanguesNaturelles` . 13
1.5 `TiLT` dans une application de vocalisation de SMS 14
1.6 `TiLT` : stratégie de correction du message SMS en "français standard" 14
1.7 Segmentation du SMS 15
1.8 Des segments aux mots 16
1.9 Arbres syntaxiques concurrents valides 23

3.1 Illustration graphique de la concordance sur le critère k avec l'assertion $h_i S h_j$ . . 54
3.2 Illustration graphique de la discordance sur le critère k avec l'assertion $h_i S h_j$ . . 55
3.3 Interprétation des relations de surclassement établies par `ELECTRE III` en vue de répondre à une problématique de classement, illustration de la prise en compte de l'incomparabilité entre h_1 et h_3. 56
3.4 Affectation des hypothèses aux classes définies *a priori* en utilisant la méthode `ELECTRE TRI` 57
3.5 Structure de préférences regroupant les relations de surclassement établies entre les hypothèses concurrentes 60
3.6 Représentation sous forme de graphe d'une structure de préférences, où un arc orienté de h_i vers h_j signifie $h_i S h_j$. L'absence d'arc dénote une situation d'incomparabilité. h_5 est surclassée par toutes les autres hypothèses. 61

4.1 Cas d'utilisation du module de contrôle décisionnel par les différents acteurs . . . 72
4.2 Identification des acteurs logiciels par abstraction des notions de module de traitement et d'hypothèse linguistique 74
4.3 Modèle du domaine du module de contrôle décisionnel 76
4.4 Vers une architecture de traitement contrôlé 79
4.5 Diagramme de séquence : Regroupement des hypothèses comparées et leurs critères associées au sein d'une structure de comparaison 81
4.6 Diagramme de séquence : Instanciation et stockage de critères associés aux hypothèses linguistiques 82
4.7 Diagramme de séquence : processus d'application d'opérateurs de surclassement pour des problématiques de classement ou de sélection 84
4.8 Diagramme de séquence : processus d'application d'un opérateur de tri par surclassement 86

4.9 Répartition des hypothèses et heuristiques d'identification des zones de préférence, d'indifférence et de veto. 95
4.10 Identification des seuils suggérés par rapport aux zones de préférence, d'indifférence et de veto 97
4.11 Un exemple de répartition particulière des hypothèses valides et non valides . . . 97
4.12 Répartition des écarts entre les hypothèses valides et non valides et identification des seuils 98
4.13 Identification du seuil optimal de coupe pour la séparation des exemples d'apprentissage positifs et négatifs 99

5.1 Processus d'analyse linguistique réalisé par TiLT pour l'indexation des requêtes . 104
5.2 Rubriques métiers associés à la requête "salon de coiffure" 105
5.3 Utilisation de la chaîne de traitement TiLT pour la résolution des chaînes de coréférence 112
5.4 Processus de transcription des SMS 132
5.5 Répartition des erreurs syntaxiques identifiées sur les corpus de LOUVAIN et du DELIC 138
5.6 Enchaînement des hypothèses générées lors du processus de transcription 141

A.1 Illustration graphique de la concordance 173
A.2 Illustration graphique de la discordance 173
A.3 Affectation des hypothèses aux classes définies *a priori* en utilisant la méthode ELECTRE TRI 176

B.1 Diagramme de classe du module décisionnel Beslissing intégré dans l'architecture de traitement TiLT 180

C.1 Interface graphique de configuration et de définition des modèles de préférences : exemple de définition des critères à utiliser 183
C.2 Interface graphique de configuration et de définition des modèles de préférences : exemple de définition d'une opération de contrôle 184
C.3 Interface graphique de construction d'un corpus de références et de la table de performances associée à partir des hypothèses concurrentes générées par un module de traitement de TiLT : application de vue (transformation XSLT sur le fichier XML des sorties) pour déterminer plus facilement leur validité 185
C.4 Interface graphique de construction d'un corpus de références et de la table de performances associée à partir des hypothèses concurrentes générées par un module de traitement de TiLT : chargement d'un fichier XML des sorties d'un module ou analyse directe d'une phrase 186
C.5 Interface graphique de construction d'un corpus de références et de la table de performances associée à partir des hypothèses concurrentes générées par un module de traitement de TiLT : intégration d'annotations dans le fichier XML des sorties . 187

D.1 Courbes de distribution construites à partir du corpus d'apprentissage pour les paires NPR-NPR sur le critère de similarité typographique 197
D.2 Courbes de distribution construites à partir du corpus d'apprentissage pour les paires NPR-NCOM sur le critère du nombre d'occurrences de l'antécédent 198
D.3 Courbes de distribution construites à partir du corpus d'apprentissage pour les paires NPR-PRON sur le critère du nombre d'occurrences de l'antécédent 198

Liste des tableaux

4.1 Tableau de performances construit à partir de l'alignement entre les sorties d'un module de traitement à contrôler et un corpus de référence approprié 91

5.1 Modèle de préférences établi par l'expert pour le classement des rubriques métiers concurrentes 108
5.2 Résultats de l'évaluation des stratégies de classement 108
5.3 Extraits des tables de performances par un alignement entre les paires candidates extraites du texte et le corpus de référence. La description des attributs et de leur type est proposée en annexe D.2. 114
5.4 Répartition des paires d'expressions extraites dans les tables de performances . . 117
5.5 Paramètres préférentiels associés par l'expert pour les paires `NPR-NPR` : extrait du modèle de préférences complet proposé en annexe D.3.1 118
5.6 Paramètres préférentiels associés par l'expert pour les paires `NPR-NCOM` : extrait du modèle de préférences complet proposé en annexe D.3.1 119
5.7 Paramètres préférentiels associés par l'expert pour les paires `NPR-PRON` : extrait du modèle de préférences complet proposé en annexe D.3.1 119
5.8 Résultats obtenus en utilisant les modèles de préférences définis *a priori* par un expert 119
5.9 Paramètres préférentiels suggérés à partir des tables de performances dédiées à l'apprentissage pour les paires `NPR-NPR` : extrait du modèle de préférences complet proposé en annexe D.3.2 120
5.10 Paramètres préférentiels suggérés à partir des tables de performances dédiées à l'apprentissage pour les paires `NPR-NCOM` : extrait du modèle de préférences complet proposé en annexe D.3.2 120
5.11 Paramètres préférentiels suggérés à partir des tables de performances dédiées à l'apprentissage pour les paires `NPR-PRON` : extrait du modèle de préférences complet proposé en annexe D.3.2 121
5.12 Résultats obtenus en utilisant les modèles de préférences suggérés par les méthodes statistiques 121
5.13 Paramètres préférentiels issus de l'alignement entre les valeurs suggérées par les heuristiques statistiques et les connaissances expertes : extrait du modèle de préférences complet proposé en annexe D.3.3 pour les paires `NPR-NPR` 122
5.14 Paramètres préférentiels issus de l'alignement entre les valeurs suggérées par les heuristiques statistiques et les connaissances expertes : extrait du modèle de préférences complet proposé en annexe D.3.3 pour les paires `NPR-NCOM` 122

5.15 Paramètres préférentiels issus de l'alignement entre les valeurs suggérées par les heuristiques statistiques et les connaissances expertes : extrait du modèle de préférences complet proposé en annexe D.3.3 pour les paires NPR-PRON 123
5.16 Résultats obtenus en utilisant les modèles de préférences mixtes 123
5.17 Résultats obtenus en utilisant les arbres de décision 124
5.18 Résultats obtenus en appliquant une méthode additive 125
5.19 Résultats obtenus en appliquant une méthode additive pondérée 126
5.20 Résultats obtenus en appliquant une méthode par vote 126
5.21 Résultats obtenus en utilisant le modèle de préférences mixte 127
5.22 Évaluation du processus initial de transcription de TiLT sur le corpus du DELIC . 135
5.23 Évaluation du processus initial de transcription de TiLT sur le corpus de Louvain 136
5.24 Paramètres préférentiels définis par un expert pour la sélection des meilleurs structures de validation et des meilleurs terminaux 147
5.25 Évaluation du processus contrôlé de transcription de TiLT sur le corpus de Louvain147

D.1 Critères disponibles et exploitées pour l'identification des cas de coréférence . . . 191
D.2 Modèle de préférences expert pour les paires NPR-NPR 192
D.3 Modèle de préférences expert pour les paires NPR-NCOM 193
D.4 Modèle de préférences expert pour les paires NPR-PRON 194
D.5 Modèle de préférences suggéré à partir des tables de performances dédiées à l'apprentissage pour les paires NPR-NPR 195
D.6 Modèle de préférences suggéré à partir des tables de performances dédiées à l'apprentissage pour les paires NPR-NCOM 196
D.7 Modèle de préférences suggéré à partir des tables de performances dédiées à l'apprentissage pour les paires NPR-PRON 197
D.8 Modèle de préférences mixte, établi par l'expert à partir du modèle de préférences suggéré pour les paires NPR-NPR 199
D.9 Modèle de préférences mixte, établi par l'expert à partir du modèle de préférences suggéré pour les paires NPR-NCOM 200
D.10 Modèle de préférences mixte, établi par l'expert à partir du modèle de préférences suggéré pour les paires NPR-PRON 201

Introduction

La démocratisation de l'informatique et des nouveaux moyens de communication comme Internet ou la téléphonie mobile entraîne une constante augmentation du nombre de documents numériques qui nous entourent. Bien que les contenus graphiques et sonores occupent une place de plus en plus importante, les données textuelles restent le type d'information le plus usuellement créé, visualisé ou manipulé.

Les ordinateurs et autres terminaux de saisie et de visualisation comme les téléphones portables constituent désormais des interfaces de communication véhiculant nos activités langagières. Initialement conçus pour traiter des langages formels, tels que les formules mathématiques ou les langages de programmation, les ordinateurs furent rapidement intégrés dans les communications entre humains, tout d'abord en tant que média d'échange et de stockage et maintenant en tant qu'interlocuteur potentiel.

Ainsi, les techniques et méthodes informatiques ont dû être adaptées pour que l'ordinateur puisse manipuler ces données issues de nos productions langagières.

Sans doute dès l'apparition des premiers ordinateurs dans les années 1950, les acteurs des recherches et développements informatiques ont tenté de reproduire et d'automatiser notre faculté de traitement des données écrites en langue naturelle[1] comme la compréhension du sens d'un énoncé écrit ou oral, la prononciation d'une succession de mots, la prise de parole cohérente dans un dialogue, le résumé ou la traduction d'un document, etc. Le Traitement Automatique des Langues Naturelles (TALN), en tant que discipline de l'intelligence artificielle, vise à reproduire ou simuler les processus cognitifs que nous mettons en œuvre lors de la réalisation de ces différentes activités. Cependant, l'apparente facilité avec laquelle nous manipulons des énoncés en langue naturelle résulte d'un long apprentissage voire de facultés innées ou tout au moins de prédispositions qui nous caractérisent et nous différencient des autres espèces vivantes.

Si le TALN constitue encore un domaine de recherche passionnant, son histoire retrace également un parcours déroutant parsemé d'échecs et de difficultés liées à l'automatisation de ces traitements complexes. La principale difficulté soulevée en une cinquantaine d'années de recherche, relève de la nature multidisciplinaire de ces recherches. En effet, le processus complet de compréhension d'énoncés en langue naturelle repose sur des connaissances complémentaires plaçant le TALN à l'intersection de plusieurs domaines tels que : la linguistique, pour la description du matériau d'analyse ; la psycholinguistique pour les aspects cognitifs ; les neurosciences pour l'organisation des composants du processus d'analyse ; l'informatique et les mathématiques pour les aspects algorithmiques et calculatoires.

[1] Terme issu de l'anglicisme "natural language", communément accepté pour désigner un système conventionnel de symboles sur lequel reposent les productions langagières des humains, qui en français serait simplement appelé langue.

Cependant, les différents travaux en TALN semblent avoir mis à l'écart une discipline pourtant très liée à nos facultés d'analyse des langues naturelles : **la décision**.

En effet, l'une des difficultés rencontrées par la plupart des systèmes de TALN est liée à l'indéterminisme des procédures d'analyse, qui se matérialise par l'apparition d'interprétations concurrentes lors des différentes tâches de traitement. L'aisance et la rapidité avec laquelle nous comprenons ou produisons des énoncés en langue naturelle sans réelle ambiguïté ou hésitation illustrent à la fois le côté artificiel de ce phénomène problématique mais également notre faculté décisionnelle. L'absence de prise en compte de ces aspects décisionnels lors de l'automatisation du processus d'analyse linguistique conduit à la génération d'interprétations concurrentes et à la propagation de ces indéterminations. Afin d'améliorer l'efficacité et la pertinence des traitements, il apparaît comme un enjeu fondamental du TALN d'être capable de contrôler la pertinence de ces interprétations concurrentes.

Le TALN s'est donc principalement focalisé sur les aspects liés à l'analyse, c'est-à-dire à l'exploitation de connaissances disponibles par les algorithmes de traitement en vue de produire des structures linguistiques. Nos travaux ont pour objectif de compléter ces procédures d'analyse par des mécanismes de décision offrant un contrôle et une évaluation des résultats générés.

En s'appuyant sur un exemple concret de système de TALN, nous cherchons à atteindre une meilleure compréhension de ce phénomène, mais également à développer un système de contrôle permettant de compléter le processus d'analyse par des phases de décision.

Considérer le contrôle d'un système de TALN comme un processus décisionnel, nous entraîne vers deux domaines de recherche jusqu'ici sans intersection : le TALN et la théorie de la décison. Afin de rendre nos travaux plus clairs pour cette dernière communauté, le `Chapitre I` sera consacré à une introduction au TALN, son orientation dans un milieu industriel et surtout la problématique à l'origine de notre projet : l'indéterminisme du processus de génération d'interprétations linguistiques.

Dans le `Chapitre II`, nous verrons que face à ce problème récurrent, de nombreuses stratégies de contrôle ont été envisagées se différenciant principalement par la nature des connaissances sur lesquelles repose la prise de décision. Nous constaterons également que, comme dans la plupart des contextes décisionnels, l'évaluation des différentes alternatives repose non pas sur un type d'information mais sur un ensemble de connaissances complémentaires.

Nous verrons alors dans le `Chapitre III` que l'Aide MultiCritère à la Décision (AMCD) offre un cadre à la fois théorique et pragmatique pour la résolution de problèmes décisionnels basés sur l'exploitation de plusieurs sources de connaissances hétérogènes et complémentaires. En s'inspirant donc des travaux issus de l'AMCD, nous avons développé et intégré dans une chaîne de traitement une méthode et un ensemble d'outils de contrôle basés sur une approche décisionnelle. Le `Chapitre IV` propose une description de ces développements en insistant notamment sur l'influence du contexte industriel dans lequel s'inscrivent nos travaux.

Le `Chapitre V` sera consacré à l'évaluation de nos travaux sur des cas concrets d'expérimentation et nous permettra de valider l'apport mutuel des deux communautés concernées, que nous avons de manière novatrice reliées autour de notre problématique.

1

Les processus de TALN et la génération des hypothèses concurrentes

Sommaire

1.1 Le traitement automatique de textes écrits 2
1.1.1 TALN : définition et enjeux 2
1.1.2 Caractériser le TALN par ses difficultés 3
1.1.3 Niveaux d'analyse 5
1.1.4 Les systèmes de TALN : entre inspirations psycholinguistiques et modèles informatiques 8
1.2 Le TALN dans un contexte industriel 10
1.2.1 Des techniques et outils d'analyse linguistique au service des applications 11
1.2.2 `TiLT` : une boîte à outils linguistique 12
1.2.3 Description détaillée d'un processus d'analyse réalisé par `TiLT` 14
1.3 Hypothèses concurrentes et ambiguïtés artificielles : un problème récurrent en TALN 17
1.3.1 Indéterminisme et génération d'erreurs d'interprétations 17
1.3.2 Quelques pistes d'explication de ce phénomène 19
1.3.3 Illustrations de la génération d'hypothèses concurrentes 20

Lors de l'introduction, nous avons situé nos travaux à l'intersection de deux domaines de recherche : le TALN et l'AMCD. Le début de ce chapitre s'adresse principalement à la communauté de ce dernier domaine, puisqu'il vise à présenter les bases du traitement d'énoncés écrits en langue naturelle. Nous nous intéresserons dans un premier temps à la description des objectifs et de la nature d'un processus d'analyse linguistique. Nous verrons ensuite que face aux différentes difficultés soulevées par l'automatisation de ce processus cognitif complexe, de nombreuses stratégies ont été envisagées.

Dans un second temps, nous nous appuierons sur un exemple concret de système de pour décrire plus en détail les différentes étapes d'un processus d'analyse. Nous constaterons également que les choix stratégiques mis en œuvre pour le développement de cette chaîne ont été fortement influencés par le contexte industriel dans lequel s'inscrivent nos travaux.

Nous introduirons finalement la problématique à l'origine de notre projet : la prise en compte et la propagation d'hypothèses concurrentes lors des différentes étapes du processus d'analyse. Après une description précise de ce phénomène problématique récurrent et de son impact sur la qualité et l'efficacité du traitement, nous énumèrerons les principales causes de son apparition. Toujours en s'appuyant sur la chaîne de traitement développée par l'équipe `Langues Naturelles` de `France Télécom division R&D`, nous illustrerons ce phénomène avec plusieurs cas concrets de traitements conduisant à la génération de multiples hypothèses d'analyse.

1.1 Le traitement automatique de textes écrits

1.1.1 TALN : définition et enjeux

La traduction automatique, l'indexation et la recherche documentaire, le dialogue homme-machine, les systèmes de correction automatique, l'aide à l'apprentissage de langues étrangères, constituent quelques exemples d'applications qui semblent à la portée des systèmes d'information(s) que nous utilisons quotidiennement. Dès l'avènement des premiers ordinateurs, les chercheurs universitaires et industriels ont identifié que la maîtrise des données numériques textuelles serait un enjeu important du XXème et du XXIème siècle. Quel constat pouvons-nous faire de ces travaux après plus de soixante années de recherche et développement ? Certains services tels que les correcteurs orthographiques et les moteurs de recherche sont utilisés quotidiennement par un grand nombre de personnes et ont atteint un très bon niveau de performance. D'autres, tels que la traduction automatique, le dialogue homme-machine ou la construction automatique de résumé ont du mal à s'imposer comme des solutions logicielles opérationnelles. Pourquoi la traduction automatique, qui fait partie des premiers projets informatiques, ne fait-elle pas partie des problèmes résolus à ce jour ? Sans prétendre pouvoir expliquer précisément les raisons de cet échec ou plutôt de ce retard, il semblerait que les difficultés engendrées par l'automatisation d'un processus de traduction aient été sous-estimées. Ces difficultés proviennent principalement de la nature du matériau d'étude : les langues naturelles. Constatant les capacités d'un ordinateur à traiter des langages formels comme les langages de programmations ou les équations mathématiques, les chercheurs ont tenté de mettre au point des automates permettant d'interpréter et de manipuler des données textuelles écrites dans une langue naturelle. Mais ceci était sans prendre en compte le caractère infini des symboles et des règles de production qui définissent une langue naturelle.

L'analyse de ces données produites lors des communications entre humains constitue le domaine d'application du Traitement Automatique des Langues Naturelles. Les travaux menés dans ce domaine ont pour objectif de fournir un ensemble de mécanismes et d'outils permettant de traiter les données produites par des communications en langue naturelle. Ces outils s'insèrent

donc dans une boucle d'interaction et de communication comprenant trois phases :

- la reconnaissance de la parole ;
- le traitement des données écrites ou transcrites ;
- la synthèse de la parole à partir du traitement effectué.

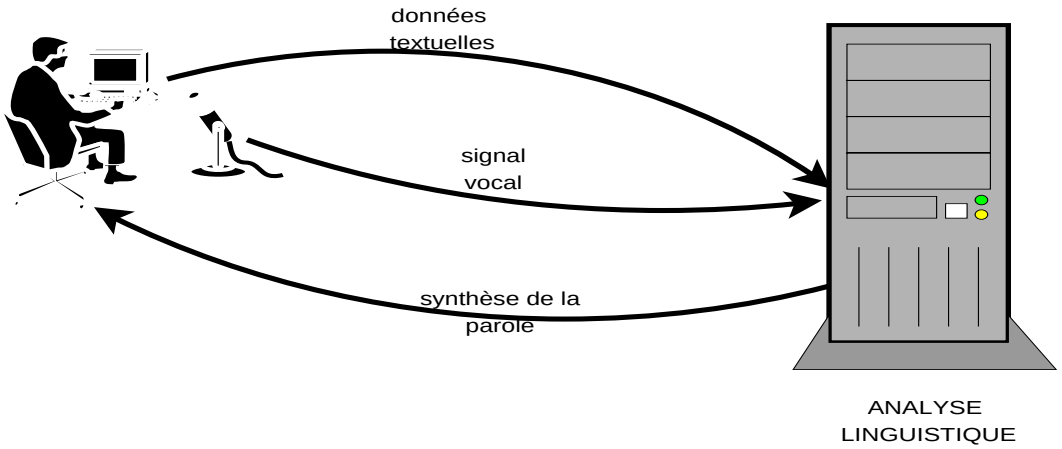

FIG. 1.1 – Le TALN : un ensemble d'outils d'interactions et de communications "artificielles"

Malgré les fortes dépendances entre ces trois étapes de simulation, la reconnaissance et la synthèse de la parole sont fréquemment considérées comme des problématiques de traitement du signal. Le TALN est donc souvent réduit à cette étape intermédiaire visant à analyser des énoncés écrits ou transcrits dans une langue naturelle. Nous conservons ainsi cette définition réductrice pour considérer le TALN comme l'ensemble des traitements linguistiques effectués pour analyser ou produire des énoncés écrits ou oraux en langue naturelle.

1.1.2 Caractériser le TALN par ses difficultés

Il serait incomplet de présenter le TALN sans introduire les difficultés induites par l'automatisation de ce processus cognitif. En effet, le demi-siècle de recherche en TALN, bien que ponctué d'échecs et de désillusions, a permis d'atteindre une très bonne compréhension des différentes difficultés rencontrées et attendues. La faculté qui caractérise sans doute le mieux les êtres humains est leur capacité à communiquer. Moyennant un apprentissage relativement intuitif, il nous apparaît relativement simple de converser ou de comprendre les propos des autres (à condition de disposer d'un bagage linguistique commun), d'apprendre une autre langue ou de produire des textes cohérents (quoique ...).

Mais le TALN, en tant que discipline de l'intelligence artificielle visant à simuler à l'aide de techniques informatiques ces capacités langagières, se heurte à de nombreuses difficultés.

La première est sans doute due à la complexité de ce phénomène, qui par nature, le place à l'intersection de nombreux domaines de recherche :

- la linguistique, pour diposer des règles et des connaissances nécessaires sur la langue traitée ;
- l'informatique, pour la formalisation de ces connaissances et la mise en place des algorithmes de traitement ;
- la neuro-science et la psycho-linguistique, pour mieux cerner l'organisation des traitements et des connaissances ;

- les mathématiques, pour la maîtrise des aspects calculatoires et l'observation statistique des phénomènes linguistiques ;
- l'ergonomie, pour rendre les communications homme-machine plus naturelles ;
- la prise de décision, objet de nos travaux ;
- etc.

En plus de ces difficultés induites par la multidisciplinarité du TALN, le matériau d'analyse, c'est-à-dire les langues naturelles, est apparu comme intrinsèquement pourvu de subtilités que les algorithmes peinent à gérer. En effet, en plus de l'impossibilité de définir en extension l'ensemble des phénomènes linguistiques envisageables dans une langue, certaines caractéristiques des actes langagiers sont extrêmements difficiles à traiter par des programmes informatiques comme l'ambiguïté généralisée des langues ou l'implicite. L'ambiguïté, qui pour les linguistes désigne une incompréhension dans une boucle d'interaction entre plusieurs locuteurs, est généralisée ici pour souligner une situation locale de "un à plusieurs" où l'analyse d'un élément engendre plusieurs hypothèses d'interprétation. Même si elles ne sont quasiment jamais perçues lors de nos communications quotidiennes, ces ambiguïtés touchent tous les éléments qui constituent un système langagier :

- les sons, un même son peut correspondre à plusieurs phonèmes possédant des fréquences proches (/i/ /u/ /y/) ;
- les mots (morphèmes), qui peuvent remplir plusieurs fonctions ("la" en pronom, nom commun ou déterminant) et avoir plusieurs sens et catégorie morpho-syntaxiques ("ferme" pouvant signifier un bâtiment, l'action de fermer, l'état de ne pas être mou, etc.) ;
- les structures de phrases, les liens entre les éléments ("Il poursuit l'homme avec le vélo.") ;
- le sens des phrases ("La belle ferme le voile.") ;
- l'intégration des phrases dans un contexte d'énonciation ("Tous nos amis ont bu un verre." ou encore la résolution des relations anaphoriques).

Ces ambiguïtés, que l'on peut certes qualifier de locales, témoignent de la quantité d'informations contextuelles que nous considérons pour interpréter les énoncés qui nous parviennent. Pour que la communication entre un homme et une machine paraisse la plus naturelle possible, il faut que la machine dispose de l'ensemble de ces informations. Il semble à l'heure actuelle utopiste d'imaginer disposer d'une modélisation des connaissances universelles et de la perception du monde que nous avons acquise en plusieurs années d'apprentissage avec une "machinerie" visiblement plus efficace que celle des machines de `Turing`. Par exemple, la compréhension d'une phrase telle que "Nicolas regarde son chien parce qu'il l'a mordu." nous apparaît comme triviale. Cependant, le processus d'interprétation que nous avons déployé pour relier "il" à "chien" et "l" à "Nicolas" repose sur une connaissance du monde, le fait qu'il soit plus probable que le "chien" ait mordu "Nicolas" plutôt que l'inverse, qui est implicite dans la phrase et évidemment difficile à modéliser et à traiter par un programme informatique.

Le TALN doit donc composer avec l'ensemble de ces difficultés pour analyser le contenu des données textuelles. L'analyse, notion quelque peu abstraite, peut donc être définie comme un processus visant à construire une représentation informatique du sens de l'énoncé traité, où énoncé correspond à une donnée de granularité variable pouvant désigner un mot, une suite de mots, une phrase, un paragraphe, un discours, un document, etc.

L'un des objectifs du TALN est orienté autour du développement d'outils et de processus permettant d'identifier et de représenter les connaissances et les structures linguistiques qui composent un énoncé formulé dans une langue naturelle. Le TALN peut ainsi être défini comme une science étudiant les moyens d'automatiser l'analyse des langues naturelles, qui constituent le matériau d'étude. Mais le TALN peut également être caractérisé par l'ensemble des techniques développées au service des applications.

1.1.3 Niveaux d'analyse

La compréhension d'un énoncé, la construction d'une représentation de son sens, repose sur différents niveaux de traitement, dont chacun constitue une spécificité de la linguistique. Cette section présente succintement l'objectif et le type d'information concernée par chacun de ces niveaux, pour une description plus détaillée le lecteur est invité à lire [Véronis, 2004], [Sabah, 1989] ou/et [Yvon, 2007].

Segmentation

La segmentation est le premier traitement effectué sur le texte à traiter. Il consiste à identifier et caractériser les éléments de base qui composent la phrase, c'est-à-dire des segments.

"Mon frère a 8 ans." ⇒ "Mon/MOT frère/MOT a/MOT 8/NOMBRE ans/MOT ./PONCTUATION"

Cette étape de l'analyse apparaît relativement simple à première vue, mais se complique largement en présence de structures plus "exotiques" comme des sigles "R.A.T.P.", des résultats sportifs "6/2-7/5-6/4" et plus généralement la reconnaissance des nombres "1.212,5" ou des noms propres "M. O'Hara".

De plus, cette étape de segmentation se complique largement lorsque la notion de mot, unité de segmentation, n'apparaît pas. C'est le cas pour certaines langues idéographiques ou agglutinantes comme le chinois et le finnois.

Certains systèmes de TALN analysent des documents textuels provenant de sources diverses. Ils doivent donc composer avec les multiples formats existants (postcript, doc, html, xml, etc.), ce qui complexifie la phase de segmentation qui doit être capable de différencier le contenu des documents, i.e. le texte à traiter, des balises et marqueurs liés au formatage.

L'analyse lexicale

L'objectif de la phase d'analyse lexicale est de transformer les segments identifiés et caractérisés lors de la phase de segmentation en "mots". Un "mot" correspond à une unité linguistique associée à des propriétés :

- phonétiques, prononciation ;
- morpho-syntaxiques, catégorie syntaxique / genre / nombre ;
- sémantiques, sens.

La technique la plus classique et la plus usitée consiste à consulter un lexique contenant l'ensemble des mots (sous leurs formes fléchies ou lemmatisées : regroupement de formes sous une même adresse lexicale) reconnus auxquels sont associées les informations citées précédemment. La puissance des ordinateurs modernes ainsi qu'une représentation optimisée de ces connaissances comme les automates à états finis permettent un accès rapide à cette ressource pouvant être très volumineuse (plusieurs centaines de milliers d'entrées).

Cette technique d'accès lexical direct est évidemment limitée dans la mesure où il est impossible de dresser une liste exhaustive de l'ensemble des mots utilisables dans une langue naturelle. De plus, les langues naturelles sont en constante évolution et de nouveaux mots (néologismes) apparaissent chaque jour ("sarkozisme", "bravitude", etc.). Valide ou non il est indispensable que les systèmes d'analyse et plus précisément le processus d'analyse lexicale soit en mesure de fournir des informations, même partielles, sur ces mots "inconnus".

Pour atteindre une robustesse satisfaisante, d'autres techniques d'analyse lexicale ont été envisagées pour générer les informations nécessaires à la poursuite de l'analyse. Ainsi, la morphologie se base sur l'analyse de composants linguistiques plus "fins" : les morphèmes, pour déduire la nature syntaxique des mots inconnus dans un lexique. Par exemple, bien que le mot "sarkozisme" soit absent des lexiques actuellement disponibles, le suffixe "-isme" caractérise un substantif masculin singulier et marque l'appartenance à un groupe ou un système. À travers l'analyse des composants du mot, l'analyse dérivationnelle garantit une meilleure robustesse au processus de traitement pour les cas d'imcomplétude du lexique. Toujours dans un souci de robustesse, l'analyse lexicale peut être complétée par des stratégies de correction afin de traiter les mots mal orthographiés. À titre d'exemple, les corrections phonétiques suggèrent de remplacer un mot mal-orthographié par ses homophones connus du lexique.

L'analyse lexicale permet également de reconnaître des découpages lexicaux complexes qui ne peuvent être identifiés lors de la segmentation. "Je mange des pommes de terre." sera reconnu lors de la segmentation comme une succession de six mots suivie d'un signe de ponctuation et un autre objectif de l'analyse lexicale est d'associer les trois segments "pommes", "de" et "terre" à la locution "pomme de terre" et ses caractéristiques syntaxiques, phonétiques et sémantiques.

L'analyse syntaxique

L'analyse lexicale décrite ci-dessus considère la phrase comme une succession linéaire de mots possédant des caractéristiques qui leur sont propres, c'est-à-dire indépendemment de leur voisinage. Cependant, toute succession de mots ne correspond pas à une phrase acceptable syntaxiquement. L'un des fonctions de l'analyse syntaxique est d'attester la validité d'une phrase en s'appuyant sur un ensemble de règles et de contraintes qui définissent une langue. La description des constructions syntaxiques valides forme une grammaire pouvant être représentée par différents formalismes.

Lors d'une première étape dite de surface (shallow parsing), l'analyse syntaxique valide localement les successions de mots en construisant des groupes de mots appelés constituants ou syntagmes, tels que les groupes nominaux, verbaux, prépositionnels, adverbiaux, etc. Quelle que soit la validité de la phrase, cette identification doit être en mesure de reconnaître les successions licites de mots "(le verre de Céline) (a été cassé) (lors du trajet)." ou "verre le (Céline) (a été cassé) de (lors du trajet).".

Cette phase d'analyse de surface est complétée par une analyse dite profonde (deep parsing) dont l'objectif est d'identifier les relations et de caractériser les fonctions qui régissent la distribution des constituants : sujet-verbe, nom-complément, verbe-complément, etc. Les relations syntaxiques identifiées entre les différents constituants forment une structure arborescente :

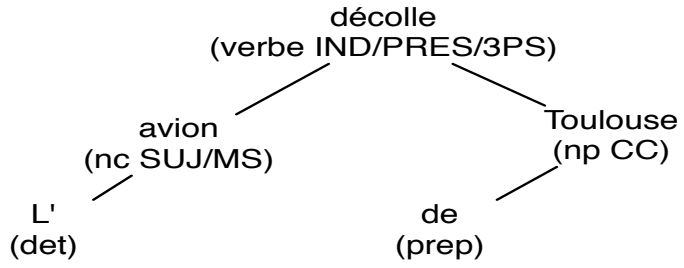

FIG. 1.2 – Représentation arborescente d'une analyse syntaxique en dépendance

Malgré l'apparente facilité que nous éprouvons pour répondre à des questions telles que : où est le verbe ? Qui fait l'action ? Qui la subit ? etc., la construction des arbres syntaxiques soulève de nombreuses difficultés (voir section 1.1.2).

L'analyse sémantique

L'analyse sémantique a pour objectif la construction d'une représentation du sens de la phrase, mais également de procéder à une vérification de sa cohérence. Bien que licite syntaxiquement, l'analyse d'une phrase telle que "Il possède le jardin de la clé du jardin" ne peut conduire à une interprétation sémantique valide.

Il convient lors de la phase d'analyse sémantique d'identifier la signification de chacun des mots de la phrase pour ensuite les mettre en relation et ainsi obtenir une structuration du sens. Ce processus d'analyse repose sur l'identification de deux sortes de connaissances sémantiques : les concepts (proposés par exemple par `Wordnet` [Miller, 1995]) qui correspondent aux objets manipulés dans le contexte d'énonciation et les structures prédicatives permettant d'attribuer des rôles aux différents concepts. Ainsi le sens d'un énoncé est fréquemment représenté par des formules logiques telles que "aimer(Jean,Sophie)", où les arguments de la structure prédicative associée au verbe "aimer" sont instanciés par les concepts "Jean" et "Sophie". Ces modèles de représentation sémantique ont cependant rapidement été confrontés à la difficulté de modéliser certains phénomènes tels que la temporalité ou la quantification.

On constate aisément que l'appariement de ces structures prédicatives sur les éléments de la phrase est fortement lié et dépendant des structures syntaxiques construites. Cependant, ce passage au niveau sémantique n'est pas qu'une simple transposition des fonctions syntaxiques aux rôles sémantiques. En effet, le sens d'une phrase n'est pas systématiquement accessible par l'analyse même de ses constituants, elle repose sur des connaissances complexes pouvant appartenir au contexte d'énonciation ou à une représentation du monde.

Malgré l'ensemble des difficultés soulevées lors de ce niveau d'analyse, les applications pouvant tirer parti de ces représentations sémantiques sont nombreuses. On peut notamment citer la traduction automatique qui a pour objectif de restituer dans une langue cible une phrase valide (syntaxiquement et sémantiquement) conservant le sens de la phrase écrite dans la langue source. De même, les systèmes de dialogue homme-machine reposent également sur la compréhension du contenu sémantique des énoncés afin de formuler des tours de parole pertinents.

L'analyse pragmatique

L'analyse pragmatique permet de situer la phrase et l'interprétation de son sens dans son contexte d'énonciation. Cette phase de projection de l'interprétation d'une phrase dans son contexte, permet notamment de lever des ambiguïtés pouvant subsister après l'analyse sémantique. Par exemple, considérée indépendamment de son contexte d'énonciation, la phrase "Il voit l'homme avec le téléscope" peut engendrer deux interprétations sémantique ; l'une où l'observateur utilise un téléscope pour voir l'homme et l'autre où l'observateur voit l'homme portant un téléscope. Ainsi, un premier objectif de l'analyse pragmatique est d'exploiter le contexte construit lors de l'analyse des phrases précédentes pour résoudre ce genre d'ambiguïtés.

Le niveau d'analyse pragmatique s'attache également à l'interprétation logique des intentions, des attitudes, des comportements et des émotions portés par les énoncés.

```
loc. A : Et si nous allions jouer au golf ?
loc. B : Il pleut !
```

Bien qu'apparaissant comme une simple constatation, l'intervention du locuteur B conduit à l'interprétation de son refus d'aller jouer au golf. Cette inférence repose sur un lien logique existant entre le fait d'aller jouer au golf et le constat d'une météo non propice à cette activité.

Le niveau pragmatique vise à traduire les intentions portées par les énoncés, mais il permet également d'analyser la structure discursive des documents, les relations rhétoriques entre les parties du discours et de procéder à une construction progressive du contexte d'énonciation.

Ce découpage d'un processus de TALN illustre les principaux niveaux d'analyse sur lesquels repose l'identification du sens des énoncés textuels. Ces niveaux sont cependant fréquemment subdivisés en sous-tâches, notamment afin d'en réduire la complexité.

1.1.4 Les systèmes de TALN : entre inspirations psycholinguistiques et modèles informatiques

Les niveaux d'analyse introduits précédemment sont fréquemment présentés comme une succession d'étapes de traitement s'insérant dans un processus séquentiel. La dépendance entre les niveaux d'analyse est cependant bien plus complexe. Il est par exemple très difficile de déterminer la catégorie morpho-syntaxique d'un mot sans s'appuyer sur les relations syntaxiques qu'il établit avec les autres éléments de la phrase ou sans considérer son rôle sémantique.

L'interdépendance et la complémentarité des différents niveaux d'analyse et de connaissance ont été mis en évidence par de nombreux travaux en psycholinguistique. L'aisance avec laquelle nous (humains) pouvons comprendre des énoncés sans hésitation apparente, semble reposer sur notre faculté à exploiter en parallèle plusieurs sources de connaissances [Altmann, 1998] [Gibson and Pearlmutter, 1998]. En considérant le TALN comme une science visant à simuler les processus humains de compréhension des énoncés en langue naturelle, de nombreux travaux se sont inspirés des modèles cognitifs issus des travaux en psycholinguistique pour concevoir des architectures informatiques de traitement [Niv, 1994] [Abney, 1989]. Ces approches, dites connexionistes, parallélisent les différents niveaux d'analyse afin d'exploiter pleinement leur complémentarité. L'objectif de ces architectures est ainsi de combiner différentes sources de connaissances et différents processus d'interprétation pour conduire le traitement plus rapidement vers une interprétation pertinente et valide par rapport aux différents niveaux d'analyse. De nombreux projets de développement d'architectures de traitement se sont inscrits dans cette approche. On peut notamment citer `CARAMEL` [Sabah, 1990] ou bien des approches basées sur des systèmes multi-agents comme [Lebarbé, 2001] et `TALISMAN` [Stefanini and Demazeau, 1995]. Cependant, bien que théoriquement justifiables, ces développements se sont en pratique heurtés à de nombreuses difficultés, telles que la gestion des communications entre les niveaux d'analyse, la formalisation des structures de connaissances ou encore l'efficacité des algorithmes de traitement.

Les travaux plus théoriques, que nous avons caractérisés par la vision scientifique qu'ils adoptent vis à vis du TALN, s'opposent ou plutôt se complètent par une approche plus pragmatique considérant le TALN comme un ensemble de techniques et d'outils au service des applications. Motivés dans un premier temps par l'intuition erronée que les langues naturelles pouvaient être traitées comme des langages formels, puis dans un second temps par le constat des difficultés induites lors de la parallélisation des niveaux de traitement, la plupart des développements se sont fondés sur des modèles informatiques "classiques" pour concevoir des architectures plus faciles à développer, à maintenir et à étendre à de nouvelles fonctionnalités.

Ainsi, pour des raisons d'efficacité et de commodité de développement, les systèmes de traitement implémentés reposent fréquemment sur un découpage du processus d'analyse en modules, où un module correspond à une unité de traitement dédiée à une tâche précise. Par exemple, chacun des niveaux d'analyse introduits précédemment est géré par un module pouvant lui-même

être subdivisé en sous-tâche afin de réduire la complexité du traitement effectué. Une stratégie d'analyse correspond donc à l'application successive de modules de traitement permettant d'atteindre progressivement le niveau d'information souhaité. Parmi les implémentations résultant de ce modèle modulaire et séquentiel, nous pouvons citer les chaînes de traitement `GATE` [Cunningham *et al.*, 2001] et `Linguastream` [Bilhaut, 2003], mais bien d'autres analyseurs plus spécialisés sur un niveau d'analyse reposent sur ce modèle [Bourigault *et al.*, 2005].

La Fig. 1.3 propose un exemple de schématisation d'un découpage modulaire du processus d'analyse. Compte tenu de la diversité des connaissances sur lesquelles repose l'analyse de phénomènes linguistiques particuliers comme la coréférence, la disposition des sous-modules peut être discutée.

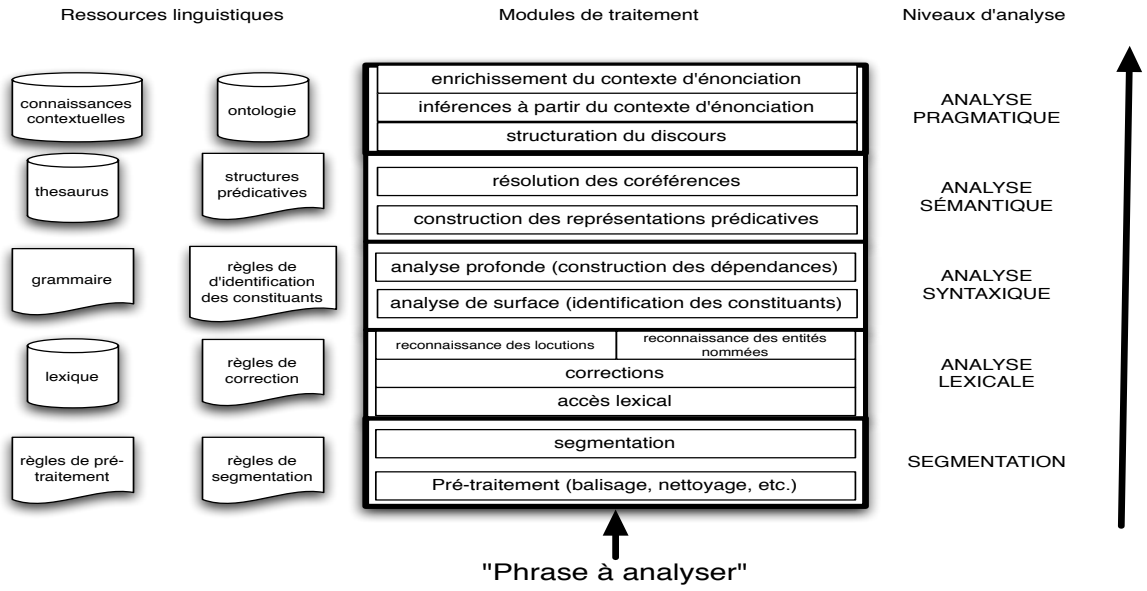

FIG. 1.3 – Modularité des processus de TALN

Bien que la majorité des travaux actuels en TALN s'orientent principalement autour du développement de processus de traitement modulaires et séquentiels, les systèmes envisagés diffèrent selon la nature et le format des connaissances linguistiques exploitées par les modules de traitement. Nous avons défini un module de traitement comme une unité de traitement caractérisée par la tâche d'analyse qui lui est dédiée et également par les ressources linguistiques qui lui sont associées.

Historiquement, la première approche proposée consistait à décrire les règles et connaissances qui définissent une langue naturelle. Ces descriptions s'appuient sur les niveaux d'analyse pour tenter d'énumérer par exemple :

- les règles de segmentation, qui permettent le découpage d'une phrase en segments ;
- les unités lexicales, c'est-à-dire les différents mots reconnus d'une langue dont l'ensemble forme un lexique ;
- les règles d'identification des constituants syntaxiques ;
- les règles de calcul des dépendances syntaxiques ;
- les concepts et règles d'interprétation sémantique ;
- etc.

Les systèmes de traitement exploitant de telles ressources descriptives sont qualifiés de symboliques et cette exploitation a été envisagée soit par l'intermédiaire de méthodes expertes soit par des méthodes issues de l'apprentissage automatique, l'apprentissage par analogie est un exemple de cette dernière méthode [et Alexandre Patry, 2007]. Chaque module qui compose ce genre de système exploite les ressources qui lui sont dédiées pour générer une interprétation de l'énoncé traité correspondant au niveau d'analyse qui lui est associé.

Constatant l'impossibilité de décrire exhaustivement les phénomènes linguistiques appartenant aux différents niveaux d'analyse et la difficulté de formaliser ces connaissances, les analyseurs symboliques ont été concurrencés par les approches statistiques [Charniak, 2000] [Collins, 2003]. Pour éviter le développement manuel fastidieux des ressources nécessaires aux approches symboliques, les analyseurs statistiques exploitent des connaissances linguistiques observées et collectées automatiquement sur de larges collections de textes annotés manuellement. L'ensemble des phénomènes observés constituent un modèle de langue qui est ensuite exploité lors de l'analyse des énoncés pour identifier les structures linguistiques connues. Les approches statistiques ont été présentées comme une alternative prometteuse aux approches symboliques, en proposant d'utiliser des descriptions linguistiques plus fines et plus spécialisées vis à vis d'un contexte applicatif donné. Cependant, cette spécialisation des connaissances repose sur la disponibilité d'un corpus exploitable, c'est-à-dire suffisamment pertinent, volumineux et fiable pour qu'un apprentissage puisse être réalisé.

Or, bien que la quantité de documents textuels disponibles, qu'ils soient bruts ou annotés, ne cesse de croître, certaines langues restent peu décrites et certains phénomènes linguistiques précis comme les sous-catégorisations ou les constructions sémantiques semblent difficilement observables.

Nous avons situé un enjeu du TALN autour de la mise en relation des différents niveaux d'analyse et de connaissances linguistiques, mais il apparaît également important d'être désormais capable de coupler les deux approches présentées précédemment. En effet, les systèmes hybrides exploitant pleinement les avantages des approches symboliques et statistiques semblent constituer la stratégie de traitement la plus prometteuse comme le montre [Frérot *et al.*, 2003] pour l'analyse syntaxique et [Biskri *et al.*, 1997] dans un contexte plus applicatif, celui de la fouille de texte.

1.2 Le TALN dans un contexte industriel

À travers une brève introduction au TALN (section 1.1.1) , nous avons mis en évidence les difficultés soulevés lors du développement de systèmes d'analyse. Pour faire face à la complexité induite par l'automatisation du processus d'analyse linguistique, différentes approches et stratégies de traitement ont été proposées (section 1.1.4) et ont données lieu à des visions à la fois antagonistes et complémentaires du TALN.

Constatant le grand nombre d'applications qui pourrait bénéficier des connaissances apportées par l'analyse linguistique sur les informations textuelles manipulées, les industriels se sont fortement intéressés aux travaux théoriques et méthodologiques réalisés en TALN. Nos travaux s'inscrivant dans un tel contexte industriel, nous allons voir dans cette section comment l'équipe `Langues Naturelles` du groupe `Orange Recherche et Développement` a abordé ce domaine pour concevoir un système complet de TALN nommé `TiLT` [Guimier De Neef *et al.*, 2002].

Dans un premier temps, nous verrons en quoi les impératifs du contexte industriel ont influencé les choix stratégiques de conception du système. Nous présenterons dans un second temps

en détail la chaîne de traitement `TiLT`, pour finir sur quelques exemples d'utilisation de ce système.

La présentation de la chaîne de traitement étudiée nous permettra également de déceler quelques pistes d'explications de la problématique à l'origine de notre projet (section 1.3) et également l'approche que nous avons envisagée pour la résoudre.

1.2.1 Des techniques et outils d'analyse linguistique au service des applications

Lors de la création de l'équipe `LanguesNaturelles` à `France Télécom` il y a une quinzaine d'années, le TALN constituait un domaine de recherche exploratoire nourrissant cependant de vifs espoirs en terme d'amélioration des services et des applications de traitement de l'information proposés.

Parmi les applications visées dans un tel contexte industriel de télécommunication, nous pouvons citer :

- les services liés à la gestion des communications : la transcription de SMS, les systèmes de dialogue homme-machine, les serveurs vocaux, etc.
- les services liés au portail Internet `Orange` : l'indexation et la recherche documentaire, la construction d'abrégés et de résumés de textes, la génération de didacticiels pour l'apprentissage de langues étrangères, etc.
- des services permettant d'éprouver les outils linguistiques développés : la traduction automatique, la construction de sources de connaissances, etc.

Pour envisager ces services basés sur une meilleur maîtrise de l'information textuelle, l'équipe `LanguesNaturelles` a développé progressivement ses propres technologies de traitement. Les ambitions en terme de développement étaient au début relativement humbles. Les premiers travaux concernaient principalement des tâches jugées désormais "simples" telles que la segmentation (pour la plupart langues non agglutinantes et non idéographiques) et l'accès à des ressources lexicales. Mais progressivement, de nouvelles fonctionnalités d'analyse et d'interprétation ont été proposées afin de compléter les traitements disponibles :

- l'analyse syntaxique de surface et profonde ;
- l'analyse sémantique ;
- la construction de représentations ontologiques ;
- la génération d'énoncés à partir d'une représentation sémantique pivot ;
- l'identification de chaînes de coréférence ;
- l'extraction de mots clés ou de thèmes à partir d'un texte ;
- etc.

Rapidement, le problème de l'interopérabilité de ces différents modules de traitement s'est posé (section 1.1.4). Cependant, dictés par des impératifs applicatifs privilégiants les aspects fonctionnels et opérationnels par rapport aux théories, les choix stratégiques en terme de conception et d'architecture de traitement s'imposèrent d'eux-mêmes. En effet, l'objectif de ces développements était de disposer à terme, d'un ensemble d'outils de traitement linguistique suffisamment génériques et adaptables pour être exploités dans différents contextes applicatifs, mais également pour être portés sur différentes langues.

Comme de nombreux systèmes industriels de TALN, les choix stratégiques se sont portés vers des modèles informatiques simples et efficaces pour obtenir des outils :

- `adaptables` aux différents types de données textuelles traitées ainsi qu'aux différentes langues ;
- `extensibles`, pour intégrer de nouveaux modules de traitement ;

- `paramétrables`, afin de simplifier l'adaptation des outils aux contextes applicatifs sans remettre en cause l'ensemble du système ;
- `efficaces`, pour être déployés sur des applications grand-public ;
- `à large couverture`, c'est-à-dire capable de couvrir le plus grand nombre possible de phénomènes linguistiques ;
- `robuste` pour faire face aux données textuelles non prévues.

1.2.2 `TiLT` : une boîte à outils linguistique

Des propriétés à l'implémentation

La recherche d'adaptabilité et d'extensibilité du système a conduit au morcellement en modules de traitement du processus d'analyse. Ainsi chacune des fonctionnalités d'analyse citées précédemment est conçue comme une unité spécifique de traitement contribuant à l'interprétation des énoncés. Pour exploiter pleinement la modularité du processus d'analyse, le choix d'architecture de traitement s'est porté sur une approche séquentielle, permettant d'appeler successivement les différents modules de traitement et ainsi d'atteindre le niveau d'interprétation souhaité.

Dans la section 1.1.4, nous avons présenté les deux principales approches envisagées en TALN pour le développement des modules d'analyse, l'une basée sur une description de la langue (symbolique), l'autre basée sur un modèle de langue acquis automatiquement sur corpus (statistique). Cependant, lors du développement des premiers modules de `TiLT`, très peu de corpus étaient disponibles, surtout dans la langue initialement étudiée, le français. L'approche symbolique a donc été privilégiée, ce qui permet d'expliquer le profil des membres de l'équipe orienté autour de deux pôles de compétences : la linguistique pour le développement de ressources décrivant les langues et l'informatique pour le développement des algorithmes de traitement.

Chaque module qui compose le système `TiLT` (Fig. 1.4) est alors constitué d'un algorithme de traitement et d'un ensemble de ressources linguistiques lui permettant d'effectuer le niveau d'analyse qui lui est dédié. L'adaptabilité des modules aux particularités du contexte d'analyse ainsi qu'aux différentes langues traitées repose sur la rigoureuse séparation qui est réalisée entre le code des algorithmes de traitement et les ressources linguistiques.

Cette approche fonctionnelle, par opposition aux modèles déclaratifs, place ainsi les ressources linguistiques au cœur de l'analyse. L'adaptation du processus de traitement aux particularités d'un contexte applicatif repose donc essentiellement sur la spécialisation, la sélection, l'ajout et la suppression de ressources linguistiques.

Un comportement dépendant des paramètres donnés par des experts

`TiLT`, de par son approche symbolique et fonctionnelle, constitue un outil de traitement hautement dépendant des ressources linguistiques qu'il utilise. La mise en place d'une stratégie d'analyse repose ainsi sur la prise en compte de l'expertise des linguistes de l'équipe, qui identifient les ressources pertinentes à exploiter pour le traitement attendu. L'intervention de ces experts sur la configuration du processus de traitement s'effectue à deux niveaux :

1. pour la mise en place d'une stratégie d'analyse ;
2. pour l'affectation des ressources linguistiques aux modules inclus dans la stratégie d'analyse.

La stratégie d'analyse définit le comportement global du système pour les différentes portions des documents analysés. Elle précise notamment quels modules doivent être activés et dans quel ordre pour chaque document, paragraphe ou phrase analysé.

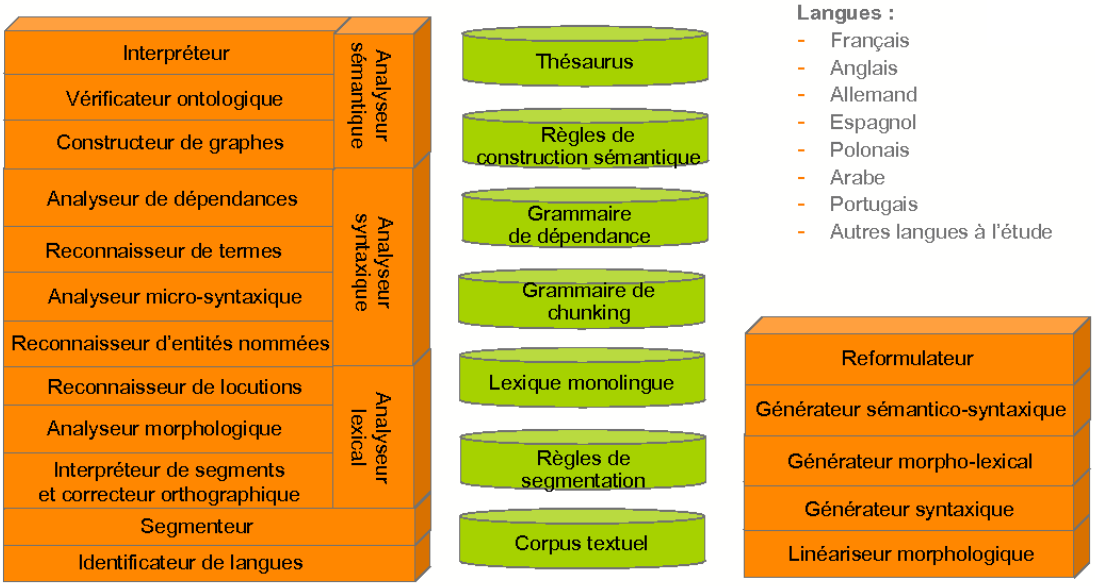

FIG. 1.4 – `TiLT` : la boîte à outils linguistique développée par l'équipe `LanguesNaturelles`

Par exemple, si l'on se place dans un contexte d'extraction d'entités nommées, la stratégie adéquate à déployer sera d'activer pour chaque phrase :

1. l'identificateur de langue ;
2. le module de segmentation en paragraphes, phrases et mots ;
3. le module d'analyse lexicale ;
4. (éventuellement des stratégies de correction et une analyse morphologique) ;
5. le module de reconnaissance de locutions ;
6. le module de reconnaissance des entités nommées.

En plus de la description de la stratégie d'analyse à adopter, le processus de traitement s'appuie sur un fichier contenant l'ensemble des paramètres techniques que nous appelons un profil d'analyse. Ce fichier de configuration décrit pour chaque module de traitement activé les ressources linguistiques qui lui sont affectées. Sans chercher à effectuer une description exhaustive des informations pouvant composer un tel profil, voici une présentation des principales sections de paramètres correspondant aux modules prépondérants du processus d'analyse :

- segmentation :
  - fichier de règles de segmentation
- analyse lexicale :
  - fichier lexique principal compilé
  - fichier lexique spécifique
  - stratégies de correction pour les mots inconnus
- Analyse syntaxique :
  - règles de construction des constituants syntaxiques
  - grammaire de dépendance utilisée
- analyse sémantique :
  - thesaurus définissant les concepts associés aux mots

– fichier de description des structures prédicatives

– ...

Outre la description de l'emplacement des ressources linguistiques affectées aux différents modules de traitement, le profil d'analyse comporte un certain nombre de paramètres plus techniques tels que le format de sortie (ASCII, XML), des bornes et limites d'application des règles pour contrôler les temps de traitement (par exemple, le nombre maximal de règles de dépendance à tester sur une phrase), des instructions de visualisation des résultats, etc.

Cette place prépondérante de l'expert lors de la mise en place d'une stratégie d'analyse confère au système une propriété intéressante d'adaptabilité et de liberté face à la spécialisation du traitement vis à vis du contexte applicatif. Mais elle contribue également à augmenter la complexité d'utilisation de `TiLT`. En effet, face à l'hétérogénéité des phénomènes linguistiques rencontrés et à la diversité des contextes d'utilisation de la chaîne de traitement, de nombreuses caractéristiques de traitement ont été externalisées en tant que paramètre technique. Déterminer le profil de paramètres techniques le plus pertinent pour une tâche donnée devient alors un travail fastidieux reposant sur l'expertise et les connaissances des membres de l'équipe.

Cette dernière caractéristique est cependant importante dans la mesure où, comme nous le verrons dans la section 3.2.1, elle a fortement influencé la façon dont nous avons traité la problématique présentée dans la section suivante.

1.2.3 Description détaillée d'un processus d'analyse réalisé par `TiLT`

Nous proposons dans cette partie de préciser la description de la chaîne de traitement étudiée en s'appuyant sur un exemple complet et concret d'analyse. Récemment, la chaîne de traitement `TiLT` a été exploitée pour proposer un service de correction de messages SMS en vue d'effectuer leur vocalisation (figure 1.5).

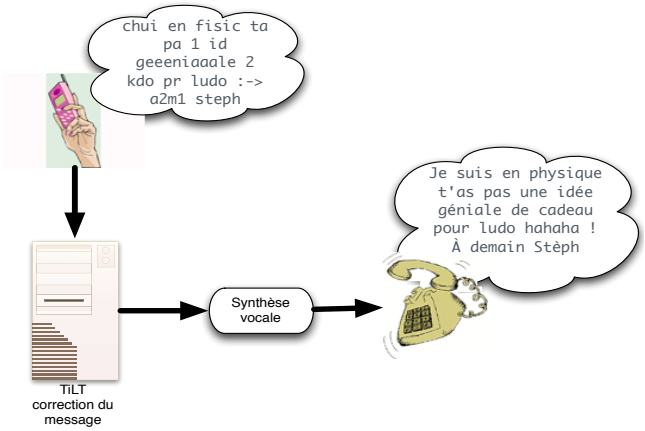

FIG. 1.5 – `TiLT` dans une application de vocalisation de SMS

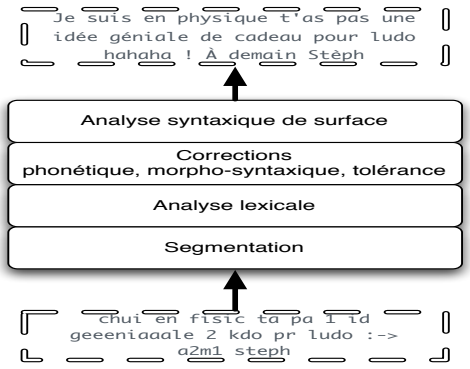

FIG. 1.6 – `TiLT` : stratégie de correction du message SMS en "français standard"

À travers cette présentation d'un cas d'usage de `TiLT`, nous souhaitons uniquement présenter la façon dont les algorithmes de traitement et les ressources linguistiques sont utilisés pour obtenir le niveau d'interprétation linguistique nécessaire à l'étape de synthèse vocale. Le lecteur intéressé pourra consulter [Guimier De Neef *et al.*, 2007] pour une description plus fine et précise de cette problématique.

Segmentation

Le premier module de traitement déclaré dans la stratégie concerne la segmentation du message SMS en segments typés. Les règles de découpage et de typage sont définies par un ensemble d'expressions régulières.

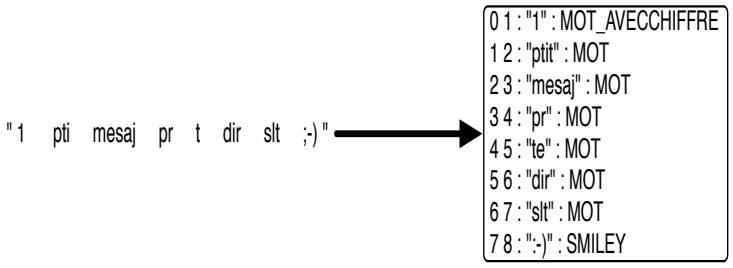

FIG. 1.7 – Segmentation du SMS

Analyse lexicale

Les segments de type `MOT` sont ensuite soumis à l'analyse lexicale. Ce module exploite un lexique de 100 000 entrées adapté au traitement des SMS, comprenant notamment une base de mots composés, des abréviations et sigles (atd = à ta disposition ; slt = salut ; etc.) et des prénoms. La phase d'analyse lexicale permet ainsi d'associer à chaque segment reconnu dans le lexique des informations telles que : lemme, orthographe standard, catégorie morpho-syntaxique, traits morpho-syntaxiques (pluriel, masculin, etc.) :

```
;-), [], SMI, -SMILEY_SEM/content, , , , ,
slt, [], salut(INTE)
salut, [saly], INTE, , salut, , , salut, ,
saluer, [salμe], V3,-1ERGROUPE-AVOIR-PASSIV/OUI-PASSPAR-PREPGOUV/PAR-TRANS, saluer_3, ,
,, ,
```

Corrections

Certains segments identifiés ne peuvent pas être directement reliés à des entrées lexicales, soit parce qu'ils sont mal orthographiés, soit parce qu'ils correspondent à une forme d'écriture particulière au SMS (par ex. les allongements : suuuuuuppppppeeeeeerrrrrr !!!!!!!). Afin d'être en mesure de traiter des messages comportant de telles particularités, des méthodes de correction sont appliquées :

- correction phonétique : un transducteur génère une phonétisation des segments contenant ou non des symboles et des chiffres, pour ensuite proposer les entrées lexicales correspondant à cette transcription phonétique.
- correction par découpage morpho-syntaxique : cette méthode de correction permet d'identifier et de découper les formes agglutinées (keske : qu'est ce que).
- tolérance : permet de standardiser des formes particulières telles que les smileys ;-), allongement (ssuupppeeeerrr : super).

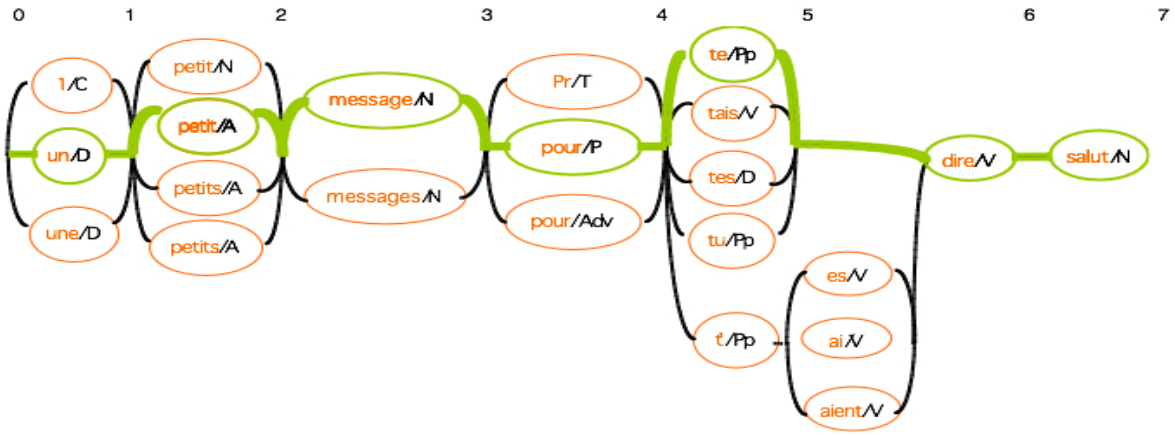

FIG. 1.8 – Des segments aux mots

Ces modes de correction complètent donc l'analyse lexicale pour garantir une certaine robustesse au processus d'analyse face à des formes d'écritures exotiques qui caractérisent les messages SMS. L'application du module d'analyse lexicale et de méthodes correctives conduit à un treillis d'hypothèses lexicales (figure 1.8). Chaque unité lexicale forme ce que nous appelons un terminal, représentant également les feuilles d'un arbre de dépendance syntaxiques.

Analyse syntaxique de surface

On constate aisément que la phase d'accès lexical et l'application de différents modes de correction entraînent la prise en compte de plusieurs formes possibles pour un segment. Bien que nous ayons largement le temps de revenir sur ce problème, l'objectif de la construction des constituants syntaxiques est d'influencer le choix d'une meilleure forme en fonction de son contexte syntaxique local. La vérification de la validité d'une forme et surtout de sa catégorie et de ses traits morpho-syntaxiques s'appuie sur une grammaire hors contexte composée d'environ 2 000 règles.

À "titre" d'exemple, la règle `GVS-CT GN-TITRE => GVS-CT-NR // EX : "c'est Mamy"` permet par exemple d'identifier une structure syntaxique récurrente dans les messages SMS correspondant à des formules de présentation ou de politesse. Pour illustrer plus clairement l'apport de cette étape, prenons le SMS suivant : "slt C mamy ca va ?". Suite à l'application des modules précédents, le segment "C" sera associé à plusieurs formes : ce, ces, ses, sais, sait, c'est, etc., l'apport de la règle présentée précédemment sera de favoriser la forme "c'est", correspondant à un groupe verbal copule (GVS-CT) car il est suivi d'une forme correspondant à un titre (GN-TITRE). L'application de cette règle conduira à la création d'un groupe jugé syntaxiquement cohérent, tout au moins localement, correspondant à un groupe verbal introduisant un nom propre (NR).

La succession de formes jugée la plus pertinente sera ensuite soumise au système de synthèse vocale afin d'être restituée au destinataire du message.

Ce cas d'utilisation de la chaîne de traitement `TiLT` nous a permis d'illustrer l'architecture modulaire et séquentielle du système, mais également la place prépondérante des ressources linguistiques dans l'analyse et notamment leur nécessaire adaptation aux particularités du contexte applicatif.

Le traitement de données textuelles issues des nouvelles formes de communication écrite comme les SMS soulève de nombreuses difficultés notamment liées aux libertés orthographiques que s'accordent les locuteurs (agglutinations : "g ésayé2tapelé" ; graphies phonétisantes et rébus "g ht du kfé a+" ; etc). À travers cet exemple et plus précisément lors de la présentation des phases d'analyse lexicale et syntaxique, nous avons également introduit un problème important : celui de la gestion des interprétations concurrentes et de l'indéterminisme.

1.3 Hypothèses concurrentes et ambiguïtés artificielles : un problème récurrent en TALN

Nous avons, lors de l'introduction sur le TALN (section 1.1.1), mis en évidence les difficultés soulevées par l'automatisation du processus d'interprétation de textes écrits en langue naturelle. Une de ces difficultés, que nous avons notamment illustrée brièvement dans la section précédente (voir Fig 1.8), concerne la multiplicité des interprétations générées par les modules de traitement. Dans cette section, nous allons dans un premier temps définir ce phénomène et poser la terminologie qui nous permettra de le caractériser. Dans un second temps, nous proposerons quelques pistes d'explication de l'apparition de ce phénomène, ce qui nous conduira à le considérer comme un problème récurrent et difficilement solvable sans exploiter des stratégies spécifiques. Finalement, nous constaterons à travers deux stratégies de traitement, que `TiLT`, comme tout système de TALN, est confronté à la prise en compte d'interprétations multiples et parfois erronées et ce, à différentes étapes du processus d'analyse.

1.3.1 Indéterminisme et génération d'erreurs d'interprétations

Les résultats d'un processus d'analyse, que nous avons appelés jusqu'ici interprétations linguistiques, reposent le plus souvent sur l'application successive et hiérarchique de modules de traitement. Au cours du processus d'analyse, la nature de ces interprétations évolue, passant de segment à unité lexicale, puis, de constituants syntaxiques à arbre de dépendance et ainsi de suite.

Nous pouvons ainsi revenir sur la définition d'un module de traitement en le présentant comme une fonction prenant en argument une interprétation proposée lors de l'étape précédente et construisant à partir des ressources linguistiques qui lui sont allouées, une nouvelle interprétation (en réalité 0, 1 ou plusieurs) correspondant au niveau d'analyse concernée. La phase d'analyse lexicale, présentée précédemment dans un contexte de correction de SMS (voir section 1.2.3), illustre cette vision générique de la notion de module de traitement. La fonction d'analyse lexicale prend en entrée un segment identifié lors de l'étape de segmentation et exploite son lexique associé pour proposer comme sortie une unité lexicale. De même, l'analyse syntaxique de surface construit à partir d'une distribution de catégories morpho-syntaxiques associées aux segments de l'énoncé, une distribution en constituants syntaxiques.

On constate cependant que ce passage d'un niveau d'interprétation à un autre s'effectue rarement de manière si déterministe en associant une sortie à une entrée. En effet, les modules de traitement sont de manière récurrente confrontés au problème de la génération de multiples sorties pour une entrée. Cet indéterminisme se matérialise par la génération d'un ensemble de sorties concurrentes parmi lesquelles certaines peuvent correspondre à des erreurs d'interprétation. Cette caractéristique de concurrence et d'incertain nous amène à introduire la notion d'**hypothèse** pour désigner toute interprétation linguistique générée par un module de traitement.

L'indéterminisme des modules de traitement peut dans certaines conditions être légitime. En effet, les différents modules qui composent un processus de TALN agissent comme des filtres successifs sur les hypothèses générées précédemment. Ainsi, certaines hypothèses locales d'interprétation, bien qu'erronées, peuvent être qualifiées de légitimes lorsque la suite du processus d'analyse est en mesure de les écarter naturellement de l'espace des hypothèses pertinentes par application de règles ou de contraintes de plus haut niveau. En effet, localement, c'est-à-dire du point de vue du module en cours d'application, cet indéterminisme est légitime dans la mesure où il ne dispose pas des connaissances nécessaires à l'évaluation de la pertinence relative des hypothèses concurrentes. Dans une approche modulaire et séquentielle, cette tâche de validation ou d'invalidation incombe aux modules ultérieurs.

Cependant, on constate que les hypothèses issues d'erreurs locales d'interprétation sont propagées lors du processus d'analyse et conduisent à des résultats finaux erronés. Au lieu d'agir comme des filtres ou des contraintes "naturelles" les modules de traitement génèrent à partir des hypothèses propagées de nouvelles interprétations. "L'explosion combinatoire" qui en résulte a évidemment un impact négatif important sur le processus d'analyse. Les traitements suivants sont alors dupliqués sur l'ensemble des branches d'analyse envisagées, augmentant ainsi les temps d'exécution et l'espace mémoire utilisé pour le stockage de ces structures de données.

La propagation des hypothèses erronées a surtout un impact important sur la qualité des résultats finaux proposés par le système d'analyse. La perfectibilité des résultats générés et surtout l'absence de validation de la pertinence des hypothèses proposées rend difficile le développement et le déploiement d'applications basées sur un processus de TALN. Quelle crédibilité accorder aux résultats d'un processus de TALN si une bonne interprétation se trouve noyée parmi un ensemble d'erreurs d'analyse ? Nous constaterons dans la section 1.3.2 que la génération d'hypothèses erronées peut être observée dès les premières phases du processus d'analyse. La propagation de ces erreurs peut conduire à des résultats totalement incongrus et parfois déconcertants. Dans [Piron, 1994], l'auteur illustre l'impact de la propagation d'hypothèses erronées sur les résultats finaux générés par un système de traduction automatique ; par exemple “In such a case, you can make a very good case for wooden cases.” est traduit en “Dans un tel cas vous pouvez faire un très bon cas pour des cas inexpressifs.” ou encore “He was sorting out food rations and chewing gum.” en “Il triait dehors rations de nourriture et mastiquant la gencive.”

Ce phénomène que nous décrivons comme lié à l'indéterminisme des processus de traitement et à la génération d'erreurs d'interprétation a souvent été rapproché de la notion d'ambiguïté [Cardey-Greenfield, 1996]. Bien que la définition proposée par le TLFI[2] corresponde au phénomène que nous venons de décrire : "Caractère de ce qui est susceptible de recevoir plusieurs interprétations", nous préférons conserver le terme d'indéterminisme pour caractériser l'absence de validation des hypothèses générées conduisant ainsi à la propagation d'erreurs d'interprétation. En effet, l'ambiguïté désignerait un état du processus où plusieurs interprétations valides (selon les règles de construction d'énoncés propres à une langue) sont envisageables, alors qu'indéterminisme caractérise bien le fait de considérer plusieurs hypothèses sans pour autant statuer sur leur pertinence ou de leur conformité.

[2]Trésor de la Langue Française Informatisée http://www.tlfi.fr

1.3.2 Quelques pistes d'explication de ce phénomène

Bien que cette problématique liée à la propagation d'hypothèses erronées résulte de l'observation du comportement particulier de la chaîne de traitement `TiLT`, il ne semble pas qu'il existe de stratégie d'analyse qui ne soit pas confrontée à ce phénomène. Nous attesterons du caractère récurrent de ce problème en montrant que pour chaque approche de traitement envisagée, des stratégies ont été proposées pour traiter ce problème (voir section 2.1).

Intuitivement, l'ambiguïté naturelle des langues a constitué la première source d'explication de l'indéterminisme des systèmes de TALN. Quelle que soit la situation de communication entre humains, cette ambiguïté est très peu perceptible. Nous avons cependant vu lors de la section 1.1.1 que l'automatisation du processus d'analyse linguistique conduisait à considérer chaque élément d'un énoncé comme potentiellement ambigu, c'est-à-dire "susceptible de recevoir plusieurs interprétations". Cet aspect fortement gênant lors de la conception d'applications informatiques constitue une caractéristique forte des langues naturelles, les opposant ainsi notamment aux langages de programmation. Cependant, l'ensemble des difficultés rencontrées ne s'explique pas uniquement par cette caractéristique du matériau analysé.

En effet, [Rady, 1983] a montré, en s'appuyant sur les systèmes d'analyse disponibles à cette époque, que l'indéterminisme des processus de traitement ne résultait pas uniquement du caractère ambigu des langues naturelles. Bien que datant de plus d'une vingtaine d'années, ces explications et conclusions restent valables de nos jours. Il apparaît en effet, comme nous l'avons souligné dans la section 1.1.4, que les choix en terme de conception des architectures de traitement et de stratégie d'analyse constituent également des pistes d'explication de l'indéterminisme des processus d'analyse. En présentant le TALN à travers sa décomposition en niveaux d'analyse, nous avons également soulevé le problème de la coopération de ces sources de traitement et de connaissances. La construction d'une interprétation, qu'elle soit d'ordre lexical, syntaxique ou sémantique, repose sur la mise en parallèle de plusieurs sources de connaissances. Or, les architectures modulaires et séquentielles ne permettent pas d'exploiter la complémentarité des différents modules de traitement. Il apparaît ainsi légitime que l'ensemble des hypothèses locales d'interprétation soient conservées jusqu'à ce que des connaissances de plus haut niveau puissent lever ces "ambiguïtés".

[Rady, 1983] (chapitre 2 section 5, 6 et 7) souligne également que les choix en terme de formalisme de représentation des règles de construction (pour la syntaxe), le choix entre une approche fonctionnelle ou déclarative ainsi que la stratégie d'analyse montante (partant des mots pour identifier la structure d'une phrase) ou descendante (partant des structures de phrases reconnues pour identifier ses composants), peuvent constituer des sources d'explication de l'indéterminisme des procédures de traitement.

Ces différents arguments offrent des explications et des pistes de réflexion quant au phénomène de la génération d'hypothèses concurrentes par un module de traitement. Ils ne permettent cependant pas d'expliquer pourquoi un bon nombre de ces hypothèses concurrentes sont erronées, ni pourquoi il est difficile d'éviter leur propagation afin de supprimer les analyses conduisant à des impasses ou à des interprétations finales totalement incorrectes.

L'indéterminisme local des systèmes de traitement et de `TiLT` en particulier, ne constitue pas un défaut en soi. Il paraît en effet légitime de considérer à un moment donné du processus plu-

sieurs hypothèses lorsque les informations nécessaires pour statuer de leur validité respective ne sont pas encore disponibles. Ainsi, considérer trois hypothèses de catégorie morpho-syntaxiques (pronom, déterminant, nom commun) pour un segment comme "la" suite à l'analyse lexicale ne constitue pas un défaut de l'approche, dans la mesure où dans une phrase telle que "Je donne le la.", l'interprétation valide du segment "la" en tant que nom commun sera sélectionnée lors de l'application du module de traitement syntaxique.

On constate cependant que l'application successive des différents modules de traitement ne constitue pas toujours ce filtre "naturel" permettant de réduire progressivement les branches incorrectes de l'arbre des analyses et de le restreindre à l'ensemble des hypothèses valides. La propagation des hypothèses erronées lors du processus s'explique donc principalement par la perfectibilité des ressources linguistiques utilisées. Dans une approche symbolique comme celle déployée par `TiLT`, les ressources linguistiques définissent les connaissances, règles et contraintes qui permettent d'attester de la conformité d'un énoncé pour une langue donnée. Très longtemps, le défi des linguistes était de développer des ressources pour décrire le plus de phénomènes linguistiques possibles. Un rapport de symétrie semble cependant apparaître entre les notions de couverture et de précision. En effet, la recherche de robustesse, c'est-à-dire être tolérant face à l'hétérogénéité et la multiplicité des structures linguistiques envisageables, entraîne une perte de précision dans la description des phénomènes traités. Cette augmentation de la couverture des différents modules de traitement se matérialise le plus souvent par l'intégration de règles plus génériques, par le relâchement ou la sous-spécification de contraintes mais également l'intégration de méthodes de correction permettant de traiter une partie des phénomènes inconnus.

Nous avons vu également que l'adaptation d'une stratégie d'analyse aux particularités d'un contexte applicatif reposait sur la spécialisation des ressources linguistiques. Ce travail d'adaptation est cependant très délicat et fastidieux, notamment dans la mesure où il est difficile de savoir *a priori* si une ressource, une règle ou une contrainte doit être conservée, complétée ou supprimée. Bien que des travaux aient été proposés pour spécifier automatiquement des ressources par rapport à un corpus représentatif d'un domaine d'analyse spécifique [Grabar and Zweigenbaum, 2005] [Lopez *et al.*, 2002], ils dépendent de la disponibilité de corpus spécialisés représentatifs [Péry-Woodley, 1995] et ne concernent que certains types de ressources.

Malgré le fait que la construction d'hypothèses locales concurrentes semble correspondre à un phénomène artificiel fortement lié à l'automatisation et sans doute à la simplification d'un processus cognitif complexe, la génération et la propagation d'interprétations erronées apparaît comme un défaut fortement préjudiciable et à un aspect perfectible des systèmes de TALN. Ainsi, après de nombreux travaux et avancées en terme de robustesse, l'un des enjeux du TALN repose sur l'évaluation de la pertinence des hypothèses générées qui devient comme nous allons le constater dans le chapitre suivant une problématique à part entière.

1.3.3 Illustrations de la génération d'hypothèses concurrentes

Afin de mieux cerner ce phénomène de la génération d'hypothèses concurrentes et parfois erronées, nous proposons dans cette partie de décrire le comportement de la chaîne de traitement `TiLT` sur deux exemples de stratégies d'analyse. Ceci nous permettra notamment d'identifier l'origine de l'indéterminisme mais surtout, de mettre en évidence l'intérêt de la mise en place de stratégies spécifiques visant à limiter la propagation des hypothèses erronées.

Correction de SMS

La stratégie d'analyse présentée dans la section 1.2.3 pour effectuer une correction des messages SMS en vue de leur vocalisation, proposait une première illustration du problème de la génération d'hypothèses concurrentes. En effet, nous avons vu que la correction d'un SMS en français "standard" nécessitait de disposer d'informations lexicales, phonétiques et morpho-syntaxiques sur les différents segments du message et que ces informations étaient obtenues par consultation d'un lexique spécialisé et par l'application de méthodes de correction et de déduction. À travers les exemples suivants, on constate que ces méthodes tendent à associer à chaque segment, non pas une unité lexicale mais un ensemble d'hypothèses concurrentes. L'accès lexical direct pour le segment "ferme" pourra correspondre à la forme fléchie de trois unités lexicales, en tant que nom commun, verbe et adjectif. De même, par correction phonétique, un segment comme "ferm" générera "ferme", "fermes", "ferment" avec différents traits morpho-syntaxiques, augmentant ainsi le nombre de candidats.

Une petite expérience sur un échantillon de 4364 messages SMS (issu du corpus développé par l'université de Louvain [Fairon and Paumier, 2006]) nous a permis de quantifier localement les conséquences de l'indéterminisme de `TiLT`. Pour un total de 43466 segments, 8 hypothèses lexicales sont associées en moyenne à chaque segment.

Malgré la désambiguïsation effectuée par l'analyse syntaxique de surface sur ce treillis d'hypothèses, on constate sur ce même corpus qu'en moyenne 4 distributions de constituants concurrentes subsistent.

Dans la section consacrée aux explications de ce phénomène, nous avons remarqué que l'augmentation de la couverture d'un traitement se traduisait fréquemment par la réduction symétrique de la précision des résultats générés. Dans ce contexte de correction de SMS, nous avons pu quantifier et valider ce phénomène en montrant que des améliorations du phonétiseur (utilisé pour la reconnaissance de certains segments à partir de leur phonétisation) avaient permis d'améliorer la couverture lexicale de 5%, mais que ceci s'était également matérialisé par la production d'un treillis d'unités lexicales plus complexe et ambigu, et par conséquent par une diminution de 7% de la précision du module d'analyse syntaxique de surface.

En outre, nous constatons que l'ensemble du corpus SMS a été transcrit manuellement en français standard, sans visiblement poser de problème de compréhension du message initial. Ceci nous laisse supposer qu'il existe pour chaque SMS au moins une transcription valide et donc l'obtention automatique de l'analyse correcte repose sur l'identification et la maîtrise des hypothèses erronées.

Construction des arbres de dépendance

Cette première illustration de la prise en compte d'hypothèses concurrentes lors d'un processus d'analyse, nous a permis d'identifier l'imprécision des ressources linguistiques comme la première cause d'indéterminisme. Nous proposons de compléter l'illustration de ce phénomène par une présentation de l'étape de construction des arbres syntaxiques de dépendance, qui a été communément identifiée comme une tâche complexe source d'erreurs d'interprétation.

L'identification des relations syntaxiques entre les différents constituants ou mots d'une

phrase, constitue une étape prépondérante d'un processus d'analyse, mais soulève également de grandes difficultés notamment liées à l'explosion du nombre d'hypothèses envisageables. La notion d'ambiguïté syntaxique, que nous préférons décrire sous les termes d'indéterminisme syntaxique, est très nettement illustrée par [Church, 1982], qui montre que des dizaines voir des centaines d'interprétations syntaxiques concurrentes peuvent être générées pour certains énoncés, notamment ceux incluant des syntagmes prépositionnels.

La stratégie d'analyse syntaxique conduite par `TiLT` s'expose également à l'apparition d'un tel nombre d'hypothèses concurrentes. À partir d'une grammaire de dépendance, l'analyseur syntaxique construit entre les différents constituants syntaxiques des relations de dépendance. L'attachement de deux constituants passe par l'application de règles définissant les conditions à remplir pour que la relation soit validée. On constate cependant que pour une distribution de constituants syntaxiques, plusieurs structures de relations de dépendance peuvent être envisagées. Pour illustrer ceci, nous avons analysé un ensemble de 394 phrases extraites d'un corpus de type oral (contenant des phénomènes liés à la transcription directe d'énoncés vocaux tels que les disfluences [Bové *et al.*, 2006]). Une grammaire dite générale du français adaptée quelque peu pour le traitement de ces phénomènes oraux génère en moyenne d'environ 9 arbres syntaxiques concurrents par phrase avec environ 3 analyses acceptables par phrase [3]. Bien que contrairement au cas des SMS plusieurs analyses syntaxiques puissent être valides pour une même phrase, on constate sans mal qu'une partie de ces hypothèses est erronée.

Il est délicat d'illustrer un tel phénomène par des exemples, ceux-ci étant toujours critiquables, nous pouvons à travers l'analyse syntaxique de la phrase : "Je vends des statues en bois de rose.", constater que la mise en place d'un ensemble pertinent de règles de grammaire est délicat. Parmi plusieurs dizaines d'arbres syntaxiques de dépendance concurrents proposés, la Fig. 1.9 présente les différentes interprétations considérées comme valides ou tout au moins acceptables.

La principale difficulté liée à l'analyse de cette phrase est la catégorisation fonctionnelle et l'attachement des syntagmes introduits par des mots outils tels que "des", "de", "en", pouvant à la fois introduire un syntagme prépositionnel ou nominal réalisant diverses fonctions syntaxiques. En se focalisant sur l'attachement de "en bois" au reste de la phrase, indépendamment du fait que "de rose" lui soit rattaché ou non, soit par reconnaissance d'une locution "bois de rose" soit par complément du nom (pouvant être également complément du verbe ou complément de "statues"), on constate que les règles suivantes s'appliquent et conduisent à des hypothèses erronées :

- attachement de "en bois de rose" à "statues" ;
  - `GN-NC « GP-NC : CIRC-MAN, ConditionsDependants (PREP_FORME/EN/SANS INTRO_CAS/MAN ...) //ex : "un voyage en vélo"`,
  - `GN-NC | GN-NT | PI « GP-NT | GN-HEURE | GP-YEAR | GP-MONTH | GP -NC : CIRC-TEMPS, ConditionsDependants (INTRO_CAS/TEMPS) //ex : "lundi en juillet"` ,
  - etc.
- attachement de "en bois de rose" à "vends".
  - `GV-PT « GP-NC | GVP-PN | GP-NOT : CIRC-MAN, ConditionsDependants (INTRO_CAS/MAN ...) //ex : "aller en train"`,

[3] À partir d'analyses de référence validées manuellement, seule la structure fonctionnelle des candidats est évaluée, ce qui explique le nombre assez élevé d'analyses acceptables par phrase.

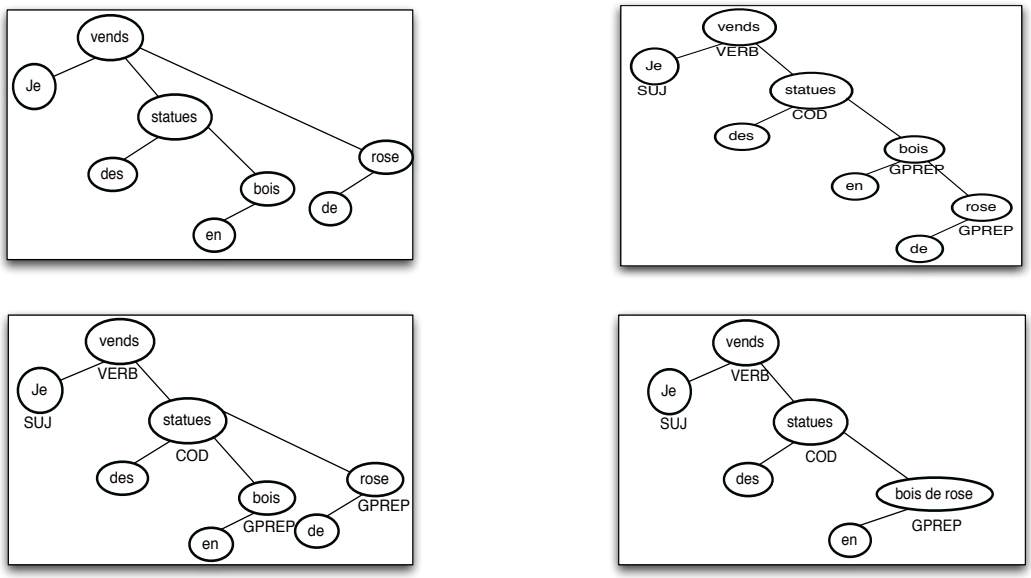

FIG. 1.9 – Arbres syntaxiques concurrents valides

- `GV-PT « GP-PQ | GP-NC : CIRC-LIEU-DEST, ConditionsDependants (INTRO_CAS/LIEU PREP_FORME/A/POUR/SUR/EN ) // ex : je pars en forêt`,
- `GV-PT « GP-PQ | GP-NC : CIRC-LIEU, ConditionsDependants (INTRO_CAS/LIEU) // ex : je pars en ville`,
- etc.

Les différentes règles illustrées précédemment montrent que les conditions d'attachement d'un syntagme prépositionnel à un verbe ou un nom reposent sur les propriétés portées par la préposition (`INTRO_CAS/LIEU, INTRO_CAS/MAN, INTRO_CAS/TEMPS`). Étant donné que dans le lexique, la préposition "en" porte tous ces traits, l'ensemble de ces règles s'appliquent et conduisent à de nombreuses structures d'arbre syntaxique. De plus, considérant que "de rose" peut être interprété comme un syntagme indépendant pouvant s'attacher à "vends", "statues" ou "bois" avec diverses fonctions, on comprend facilement l'origine de l'indéterminisme du module d'analyse syntaxique.

Basé uniquement sur des informations syntaxiques, il semble donc difficile de réduire l'espace des arbres candidats générés. En admettant que des caractéristiques syntaxiques discriminantes permettraient de spécifier les contraintes des règles d'attachement, l'analyse perdrait tout de même certainement en robustesse, surtout sur des textes dits tout-venant.

Malgré les inhérentes faiblesses de l'outil d'analyse syntaxique proposé par `TiLT`, les résultats générés sont tout de même très acceptables si l'on compare les résultats obtenus lors de la campagne d'évaluation `EASY` [Paroubek *et al.*, 2005] avec les stratégies d'analyse des autres participants.

On constate donc qu'intrinsèquement, il est difficile de réduire la propagation des hypothèses syntaxiques erronées. De plus, si l'on considère l'aspect séquentiel du processus d'analyse réalisé par `TiLT`, où les arbres syntaxiques servent à construire les graphes sémantiques représentant le sens de l'énoncé, on imagine aisément les pertes d'efficacité et de fiabilité des interprétations

générées. En effet, les descriptions des constructions sémantiques (prédicats) constituent des ressources également très difficiles à contraindre. Ainsi, à partir d'un arbre de dépendance, plusieurs dizaines de graphes sémantiques peuvent être générés.

À travers ces deux exemples de stratégies d'analyse, nous avons pu décrire et quantifier l'apparition d'hypothèses concurrentes et potentiellement d'erreurs d'interprétation lors du processus d'analyse. Nous pouvons constater que l'imprécision des ressources linguistiques utilisées par les modules de traitement constitue la première source d'erreur. Bien que des travaux très intéressants tentent de simplifier le processus d'optimisation des ressources linguistiques utilisées [Sagot and de la Clergerie, 2006], la correction de ces immenses collections de données reste principalement manuelle et constitue donc une tâche fastidieuse et délicate. Il apparaît donc indispensable de développer des stratégies spécifiques pour contrôler la génération des hypothèses concurrentes.

Un problème d'Intelligence Artificielle (IA)

En tant que systèmes basés sur l'exploitation d'informations formelles en vue de les considérer comme des connaissances, les travaux en TALN s'inscrivent par nature dans le domaine de l'IA. La maîtrise des connaissances que les humains exploitent pour effectuer des activités cognitives constitue le problème majeure du TALN et de l'IA en général. Face à la quantité et à l'hétérogénéité des connaissances à formaliser et à intégrer dans ces processus de TALN, l'apparition de situations d'indétermination devient alors légitime et comme le souligne [Bachimont, 1992], ce phénomène est une caractéristique même des systèmes d'IA.

Nous verrons au cours du chapitre suivant que ce phénomène n'est pas essentiellement artificiel, mais qu'il provient également du fait que les concepteurs de systèmes d'IA visant à simuler nos capacités cognitives, oublient une caractéristique propre à l'être humain, le distinguant notamment des autres êtres vivants, celle de disposer d'une capacité de décision.

2

Vers une approche décisionnelle de contrôle basée sur plusieurs critères de comparaison

Sommaire

2.1 Contrôler les hypothèses générées 26
2.1.1 Définition du contrôle : évaluer puis décider 26
2.1.2 Objectifs du contrôle 27
2.1.3 Des “points d’embarras” aux “points de décision” 29
2.2 Évaluer la pertinence des hypothèses à partir d’informations distinctives complémentaires 30
2.2.1 Différencier les hypothèses par observation de corpus 30
2.2.2 Différencier les hypothèses à l’aide de sources de connaissances supplémentaires ou d’heuristiques 31
2.2.3 Les "éléments de décision" dans TiLT 32
2.2.4 Des "éléments de décision" aux critères de comparaison 34
2.3 Une meilleure exploitation des critères de comparaison 35
2.3.1 Limite des approches monocritère de contrôle 36
2.3.2 Limites des approches multicritère de contrôle 37
2.3.3 Décision ou aide à la décision : vers plus de généricité et de flexibilité . 40

À travers une présentation et une description de l'indéterminisme des modules de traitement, nous avons constaté que l'obtention d'un ensemble le plus réduit possible d'hypothèses pertinentes repose sur la mise en place de stratégies spécifiques de contrôle. Principalement focalisés sur les aspects descriptifs et génératifs des structures linguistiques, les travaux en TALN ne se sont pas intéressés à cette problématique de manière systématique, alors que le contrôle des processus d'Intelligence Artificielle (IA) apparaît comme incontournable. Ce chapitre et notamment la section 2.1 visent à décrire cette notion de contrôle et à la définir sous une perspective décisionnelle.

Bien que le contrôle des hypothèses générées ne semble pas être considéré comme une des étapes légitimes des processus de TALN, nous verrons dans la section 2.2 que l'ensemble des stratégies d'analyse envisagées a été confronté à cette problématique.

Malgré l'apparente diversité des approches proposées pour traiter une source précise d'indéterminisme, nous verrons que les différentes méthodes de contrôle envisagées peuvent être formalisées sous une notion commune, celle de critère de comparaison. Cette abstraction nous permettra à la fois de positionner notre approche mais également d'introduire notre objectif : disposer d'une méthodologie et d'un ensemble d'outils génériques permettant de contrôler à l'aide des critères de comparaison disponibles les différentes sources d'indétermination pouvant apparaître au cours d'un processus de TALN .

2.1 Contrôler les hypothèses générées

2.1.1 Définition du contrôle : évaluer puis décider

Nous avons pu constater dans le Chap. 1 que la principale source de l'indéterminisme et des erreurs d'interprétation des procédures de TALN provenait de l'imprécision des ressources linguistiques utilisées. Nous avons également mis en exergue la difficulté induite par la spécialisation et l'adaptation de ces ressources aux particularités d'un contexte applicatif. Il semble également que pour des contextes d'analyse non spécialisés ou pour l'analyse de productions (écrites ou orales) spontanées, il soit impossible de contraindre l'utilisation des ressources linguistiques pour générer uniquement des interprétations fiables et cohérentes. La génération d'hypothèses concurrentes et d'erreurs d'interprétation devient alors un compromis nécessaire difficile à pallier par l'application successive de modules de traitement.

Il apparaît alors indispensable de vérifier, par l'intermédiaire de stratégies spécifiques, la pertinence des hypothèses générées au cours du processus d'analyse. L'objectif de ces stratégies est d'identifier les hypothèses les moins pertinentes, c'est-à-dire pouvant correspondre à des erreurs d'interprétation puis d'éviter leur propagation lors du processus d'analyse. Ainsi, sous le terme générique de **contrôle**, nous englobons l'ensemble des stratégies et techniques visant à évaluer la pertinence relative des hypothèses en concurrence et à effectuer en conséquence une action conduisant à écarter les interprétations erronées de l'espace des hypothèses générées. Bien que ce terme de contrôle, exploité notamment dans [Blache and Rauzy, 2006] et [Paiva and Evans, 2005], ne soit pas communément utilisé pour désigner les techniques mises en œuvre pour pallier ce problème, sa signification de mécanisme de vérification (voir le `Trésor de la Langue Française Informatisée` http://atilf.atilf.fr) définit parfaitement notre problématique. Ainsi, sous le terme de contrôle nous regroupons un ensemble de pratiques existantes, qui malgré l'absence d'une terminologie commune, visent toutes à vérifier la pertinence d'hypothèses concurrentes en vue d'identifier les erreurs d'interprétation.

La définition du contrôle que nous venons d'esquisser en nous appuyant sur notre contexte précis de TALN reprend évidemment les caractéristiques de la définition de cette tâche pour un contexte plus large, celui du contrôle des processus d'IA, notamment celle de Hayes [Erman *et al.*, 1980] :

"Le problème du contrôle est : laquelle de ses possibles un système d'IA doit-il accomplir à un instant donné du processus de résolution ?"

On constate alors que la résolution d'un problème d'IA repose sur la résolution de son contrôle. Sous le terme d'hypothèse et non d'action, la résolution d'un problème de TALN devient donc de déterminer parmi les hypothèses intermédiaires concurrentes celles qui contribueront à la construction de la solution finale.

Ces similitudes entre notre problématique de contrôle des processus de TALN et celle plus générale du contrôle en IA placent les travaux de Bruno Bachimont [Bachimont, 1992] comme une base théorique et méthodologique intéressante. En reprenant une définition proposée dans cet ouvrage page 145, nous pouvons définir les trois questions auxquelles nous allons chercher à répondre :

- "Quelles sont les connaissances de contrôle ?" ;
- "Comment les donner au système (de contrôle) ?" ;
- "Comment doit-on les utiliser ?".

2.1.2 Objectifs du contrôle

Au regard des différents travaux visant à contrôler des processus d'analyse linguistique, nous pouvons décliner l'objectif concret d'une stratégie de contrôle en trois actions différentes. En effet, en fonction du contexte d'application le rôle d'une stratégie de contrôle peut être d'effectuer une extraction des N-meilleures hypothèses, d'identifier *une meilleure* hypothèse ou bien de construire des regroupements d'hypothèses ordonnés ou non par leur degré de pertinence.

L'extraction des N-meilleures hypothèses (N-Best hypotheses)

Très largement exploitée en reconnaissance de la parole [Schwartz and Austin, 1991] [Pusateri and Thong, 2001], cette stratégie de contrôle construit un classement des hypothèses générées et exploite cette évaluation pour ne conserver que les N-meilleures.

Cette problématique de classement (ranking) ou de reclassement (reranking) a également été utilisée pour le contrôle d'autres tâches de traitement telles que l'analyse syntaxique [Collins and Koo, 2005] [Shen and Joshi, 2003] ou la traduction automatique [Callison-Burch and Flournoy, 2001] avec notamment des méthodes issues de l'apprentissage automatique telles que le `boosting`, le `perceptron`, les `SVM`, etc.

Comme nous allons le voir dans la section 2.2.1, ces stratégies s'appuient principalement sur des distributions de probabilités observées sur un corpus pour qualifier la fréquence et donc potentiellement la pertinence des différentes hypothèses considérées.

Le regroupement d'hypothèses par niveau de pertinence

Pour certains cas de contrôle, il est délicat et pas forcément souhaitable d'obtenir un classement complet et précis des hypothèses concurrentes. C'est le cas notamment des tâches de

traitement où plusieurs solutions sont acceptables comme en traduction automatique [Shen *et al.*, 2004].

L'objectif d'une stratégie de contrôle est alors d'établir une classification ou un tri des hypothèses concurrentes en fonction de leur niveau de pertinence par rapport à un besoin ou une mesure d'évaluation quelconque. Cette problématique de classification est par exemple exploitée par [Weissenbacher and Nazarenko, 2007a] à l'aide de classifieurs bayésiens pour identifier parmi l'ensemble des couples candidats `antécédent/reprise anaphorique` extraits d'un texte, ceux correspondant réellement à un cas de coréférence.

La sélection "*d'une meilleure* hypothèse"

Certaines étapes de traitement ou certains contextes applicatifs nécessitent la sélection d'une hypothèse parmi l'ensemble des interprétations candidates. Nous préférons parler de sélection *d'une meilleure* hypothèse pour conserver la possibilité que plusieurs hypothèses concurrentes peuvent être jugées comme valides ou acceptables et que la notion d'hypothèse optimale soit peu pertinente pour certaines étapes de traitement.

Cette problématique de sélection, qui revient également à effectuer un filtrage des hypothèses erronées, intervient fréquemment lorsque les résultats d'un processus de TALN rentrent en interaction directe avec un utilisateur. C'est le cas notamment lorsqu'*une meilleure* traduction doit être proposée [Akiba *et al.*, 2002], que le sens le plus pertinent d'un mot en fonction de son contexte doit être privilégié [Ide and Véronis, 1998] ou bien encore qu'une réponse pertinente doit être choisie pour un système de question-réponse [Robertson and Walker, 2001].

Ainsi, une stratégie de contrôle telle que nous la définissons revient à évaluer la pertinence relative des différentes hypothèses concurrentes et à exploiter ces jugements pour répondre à une problématique visant à identifier les erreurs d'interprétation et à éviter leur présence dans les résultats finaux proposés par le système. Cette problématique, déterminée par les impératifs du contexte applicatif, se matérialise par une des actions suivantes : la sélection, le classement ou le tri. Une stratégie de contrôle par sélection vise à extraire d'un ensemble d'hypothèses concurrentes celles qui apparaissent comme les plus pertinentes. Une stratégie de contrôle par classement produit sur l'ensemble des hypothèses comparées une relation d'ordre permettant de les ranger de la plus pertinente à la moins pertinente. Le tri, tel que nous le définissons[4], correspond à un regroupement d'hypothèses selon leur degré de pertinence. Il s'agit donc d'un ensemble de classes d'hypothèses où chacune d'elle est caractérisée par un certain degré de précision. Il apparaît important d'apporter cette précision dans la mesure où la notion de tri a un sens différent en informatique. En effet, le tri en informatique correspond à la mise en place d'un algorithme visant à ranger des éléments selon une relation d'ordre, ce qui correspond dans notre terminologie empruntée à l'AMCD à un classement.

Comme nous le verrons dans le chapitre 5, l'apport d'une stratégie de contrôle peut être quantifié à l'aide d'une mesure de précision. Cette mesure est communément utilisée pour évaluer la proportion de résultats erronés parmi l'ensemble des résultats proposés :

$$\text{précision} = \frac{\text{nombre d'hypothèses valides proposées}}{\text{nombre d'hypothèses proposées}}$$

[4] Définition adoptée dans le domaine de l'Aide MultiCritère à la Décision (AMCD)

L'ajout de points de contrôle au cours du processus d'analyse a donc pour principal objectif d'améliorer le système sur cette mesure. On constate à nouveau le rapport de symétrie que nous avons évoqué lors de la section 1.3.2 entre stratégies robustes et stratégies de contrôle, où la recherche de robustesse à travers notamment le développement de larges ressources linguistiques vise à augmenter la mesure de rappel souvent au détriment de la précision :

$$\text{rappel} = \frac{\text{nombre d'hypothèses valides proposées}}{\text{nombre d'hypothèses valides possibles}}$$

2.1.3 Des "points d'embarras" aux "points de décision"

Nous considérons donc les stratégies de contrôle comme des étapes spécifiques venant compléter le processus classique de génération d'interprétations. À travers une présentation des différentes manifestations que pouvaient prendre ces méthodes (classement, tri ou sélection), nous nous intéressons désormais à la manière dont la pertinence des hypothèses candidates peut être évaluée.

Intrinsèquement, les hypothèses générées par un module de traitement apparaissent comme un ensemble de structures de données équivalentes. Cette absence d'informations distinctives, nécessaires pour l'évaluation de la pertinence des différentes hypothèses, entraîne la propagation de toutes les interprétations générées. Ce comportement indéterministe est qualifié dans [Sabah, 1989] (Chap. 2 p. 46) de "point d'embarras". Ce terme désigne parfaitement un état du processus d'analyse où l'application d'un module de traitement entraîne la prise en compte d'hypothèses incomparables due au manque d'informations distinctives et de stratégie de contrôle.

Pour continuer sur cette terminologie empruntée à [Sabah, 1989], l'enjeu du contrôle repose sur la transition d'un "point d'embarras" à un "point de décision", qui se différencient par l'existence d'informations distinctives appelées "éléments de décision", qui sont nécessaires à la mise en œuvre d'une procédure de décision et à la réduction de la combinatoire [Vergne, 2001]. Constatant que le processus classique de génération d'interprétations linguistiques n'intègre pas ces aspects décisionnels pourtant nécessaires pour garantir la pertinence de ses résultats, nous proposons de considérer le contrôle comme un processus de décision.

Les travaux en théorie de la décision [Hansson, 1994] ont conduit à la description des étapes qui composent un processus décisionnel. Bien que l'opposition que nous avons signalée dans le section 1.1.4 entre les modèles séquentiels et les modèles parallèles ait également animé les débats dans le domaine de la modélisation d'un processus de décision, la proposition de [Mintzberg *et al.*, 1976] est la plus communément acceptée. Elle repose sur trois phases :

l'identification - générer Cette phase délimite le problème rencontré et génère les différentes solutions possibles, ce qui correspond dans notre cas à la génération des hypothèses concurrentes effectuée par un module de traitement.

le développement - évaluer L'objectif de cette phase est d'obtenir une meilleure compréhension des différentes alternatives identifiées lors de la phase précédente. Elle consiste dans notre approche au passage des "points d'embarras" aux "points de décision" par l'intégration "d'éléments de décision".

la sélection - décider Vue comme l'élaboration d'une décision optimale, cette phase exploite les évaluations de la phase précédente pour conduire à l'identification des alternatives optimales et sub-optimales. Nous rapprochons cette phase de l'élaboration d'un classement, d'une sélection ou d'un tri.

Ainsi, dans un processus décisionnel complet tel que celui esquissé précédemment notre intervention se focalise sur les phases de développement et de sélection. Le développement, qui se caractérise donc par la recherche d'une meilleure compréhension des hypothèses candidates à évaluer, repose sur l'intégration ou la prise en compte d'informations distinctives supplémentaires, ce qui constitue l'objet de la section suivante section 2.2. Nous verrons dans le chapitre suivant Chap. 3, comment notre approche propose de traiter la phase de sélection, notamment pour qu'elle puisse répondre à toutes les problématiques envisagées : le classement, le tri ou la sélection.

2.2 Évaluer la pertinence des hypothèses à partir d'informations distinctives complémentaires

Les procédures de contrôle telles que nous les avons décrites précédemment, exploitent des informations supplémentaires (non disponibles lors du processus d'analyse) en tant "qu'éléments de décision" pour évaluer la pertinence de hypothèses concurrentes. Cette section est consacrée à une présentation des différentes sources de connaissances utilisées pour générer ces "éléments de décision". Nous constaterons que quels que soient leur nature et leur type, ces connaissances peuvent être considérées comme des critères de comparaison apportant un certain point de vue de jugement sur la pertinence des hypothèses.

2.2.1 Différencier les hypothèses par observation de corpus

Lors de la section 1.1.4, nous avons introduit les deux principales approches envisagées en TALN ; l'une généralement statistique basée sur l'exploitation de corpus d'apprentissages et l'autre symbolique basée sur des ressources descriptives des phénomènes linguistiques souvent développées manuellement. Le modèle de langue issu d'un apprentissage automatique sur corpus est constitué d'un ensemble de règles hiérarchisées généralement par leur fréquence d'apparition. Ainsi, les hypothèses construites par ce genre d'approches, telles que [Charniak, 2000] et [Collins, 1997] pour l'analyse syntaxique, peuvent être évaluées à partir des fréquences liées aux règles activées. Sans exploiter d'autres ressources que le modèle de langue, les hypothèses envisagées lors du processus d'analyse statistique sont classées selon la fréquence des structures qui les composent. Le comportement déterministe de ce genre d'approche, c'est-à-dire à leur capacité à identifier "une meilleure hypohèse" parmi toutes celles générées repose sur l'exploitation de cette information de fréquence.

Cependant ce choix basé uniquement sur l'élément de décision que constitue la fréquence d'observation des structures linguistiques, ne semble pas être complètement efficace pour de nombreuses constructions de phrases incluant des phénomènes minoritairement représentés dans le corpus d'apprentissage. En effet, les techniques de programmation dynamique, largement exploitées par les systèmes probabilistes, effectuent une succession de décisions locales conduisant à la génération d'une meilleure hypothèse. Constatant que cette stratégie déterministe écartait en cours d'analyse des interprétations valides au profit d'une hypothèse erronée, des modifications ont été apportées à ces systèmes pour qu'ils soient en mesure de générer une sortie composée de plusieurs hypothèses en concurrence [Collins, 2003] [Charniak, 2005].

Ainsi ramenées à notre problème initial, les approches statistiques se sont vues confrontées à la nécessité de différentier ces hypothèses concurrentes. L'évaluation en vue d'un reclassement de ces hypothèses, technique plus souvent désigné par le terme anglo-saxon de `discriminative`

`reranking`, repose sur l'intégration "d'éléments de décision" supplémentaires comme le montre [Charniak, 2005]. Empiriquement sélectionnées, des configurations lexicales ou syntaxiques (par exemple l'attachement entre un verbe et un nom commun) sont choisies pour constituer des "éléments de décision" dont la fréquence d'apparition est observée sur le corpus d'apprentissage.

Les approches symboliques, quant à elles, ne disposent généralement pas de ces informations de fréquence permettant d'évaluer la plausibilité des hypothèses générées. On remarque cependant que sous réserve de disponibilité de corpus d'apprentissage pertinents et suffisamment volumineux, d'intéressants "éléments de décision" peuvent être extraits. Afin d'exploiter ce genre d'informations, des systèmes de traitement symboliques ont intégré ces connaissances pour évaluer la fréquence et donc potentiellement la pertinence des hypothèses générées. Par exemple, pour contrôler le processus d'application d'une grammaire de propriétés, [Blache and Rauzy, 2006] propose d'évaluer la probabilité d'apparition des patrons syntaxiques (succession de catégories morpho-syntaxiques) pour prioriser les hypothèses générées.

Ces approches hybrides ont montré dans de nombreux contextes applicatifs leur efficacité à exploiter les avantages des deux technologies, à savoir la robustesse des méthodes symboliques et l'efficacité calculatoire des méthodes statistiques. En effet, que ce soit pour l'extraction de connaissances à partir de textes [Biskri and Delisle, 1999] [Claveau, 2003] [Toussaint *et al.*, 1998], pour l'identification du sens d'un mot composé [Copestake and Lascarides, 1997], pour l'analyse syntaxique de surface [Tzoukermann *et al.*, 1995] ou pour une tâche très précise de récupération du sens d'un énoncé à partir d'analyses syntaxiques partielles [Rosé and Waibel, 1994], la combinaison d'informations symboliques et statistiques apparaît comme une approche de traitement prometteuse.

2.2.2 Différencier les hypothèses à l'aide de sources de connaissances supplémentaires ou d'heuristiques

Dans la section consacrée à l'explication de l'indéterminisme des processus de traitement et à l'identification des principales causes de ce phénomène, nous avons constaté que les systèmes de TALN, et notamment `TiLT`, ne pouvaient, de par leur modularité et la séquentialité de leur architecture, exploiter la complémentarité des différents niveaux de traitement. En effet, un module de traitement dispose généralement uniquement des ressources linguistiques décrivant la tâche qui lui incombe.

Pour pallier cette limite, des ressources linguistiques supplémentaires peuvent être intégrées sous formes de règles ou de contraintes pour compléter et surtout contraindre l'application d'un module de traitement. Cette stratégie a notamment été largement exploitée pour contrôler l'application du module d'analyse syntaxique par l'intégration de ressources sémantiques. Ces deux niveaux sont en effet étroitement corrélés et de nombreux travaux ont montré l'influence et l'apport mutuel de ces deux traitements. [Yates and Schoenmackers, 2006] propose par exemple d'intégrer des connaissances sémantiques pour filtrer les analyses syntaxiques générées. [Bourigault and Frérot, 2004] constitue un second exemple d'intégration de ressources linguistiques supplémentaires, ici des cadres de sous-catégorisation, pour contrôler l'attachement de syntagmes prépositionnels lors de l'analyse syntaxique.

L'information apportée par l'intégration de ressources linguistiques complémentaires est souvent exploitée dans une problématique de validation ou symétriquement de filtrage. "L'élément de décision" apporté par la prise en compte de ces connaissances devient alors une information distinctive de nature `binaire`, c'est-à-dire qu'une hypothèse est jugée conforme ou non vis-à-vis de ces filtres supplémentaires.

De manière beaucoup plus empirique, la pertinence des hypothèses concurrentes générées par un module de traitement peut être évaluée à l'aide d'heuristiques, traduisant la plupart du temps les intuitions ou les observations d'un expert[5] sur le matériau analysé. Les exemples d'usage d'heuristiques pour évaluer la pertinence des différentes hypothèses candidates à un moment donné du processus d'analyse sont nombreux. Pour illustrer cette stratégie de contrôle, nous pouvons citer le cas de l'évaluation des analyses syntaxiques concurrentes qui repose fréquemment sur la vérification de la conformité de ces hypothèses vis-à-vis de caractéristiques relevant davantage d'intuitions que de propriétés. La difficulté soulevée par l'attachement d'un syntagme prépositionnel au reste de la phrase a donné lieu à la formulation de nombreuses heuristiques [Whittemore and Ferrara, 1990] telles que : l'attachement droit, qui consiste à favoriser l'attachement d'un nouveau syntagme à la droite de son syntagme adjacent gauche ; l'attachement minimal, visant à préférer les arbres syntaxiques intégrant le moins de nœuds ; préférer certaines formes récurrentes d'arbres ; etc.

Dans un autre contexte de contrôle, celui de la sélection d'un antécédent pour un pronom parmi plusieurs candidats, [Mitkov, 1998] a montré que la résolution d'anaphores pronominales basée sur l'exploitation de propriétés simples, telles que tester la présence de certaines marques lexicales ou morpho-syntaxiques, la fréquence d'apparition de l'antécédent candidat dans le texte, etc., conduisait à des résultats parfois aussi bons qu'en utilisant de larges ressources linguistiques.

L'évaluation des hypothèses concurrentes sur laquelle repose la mise en place d'une action de contrôle (sélection, classement ou tri) s'appuie donc fréquemment sur l'intégration de méthodes heuristiques, qui permettent, à faible coût, de disposer des "éléments de décision" nécessaires au passage des "points d'embarras" aux "points de décision". Si un des objectifs d'une stratégie de contrôle est d'exploiter ces informations distinctives pour augmenter la précision des résultats générés, il apparaît également important et intéressant d'évaluer l'efficacité de ces différentes sources de jugement pour une tâche de contrôle donnée.

2.2.3 Les "éléments de décision" dans `TiLT`

Constatant l'inefficacité ou plutôt le manque de précision de certains des modules qui interviennent dans le processus d'analyse conduit par `TiLT`, des "éléments de décision" tels que nous venons de les présenter, ont été intégrés pour évaluer les hypothèses concurrentes générées. Comme nous allons le constater sur plusieurs exemples, les procédures *ad hoc* intégrées dans `TiLT` pour générer des "éléments de décision" exploitent également des sources de connaissances telles que : des observations sur corpus, des ressources linguistiques complémentaires ou des connaissances plus empiriques (heuristiques, intuitions).

Contrôle lexical

L'application du module d'analyse lexicale sur les résultats du module de segmentation conduit à la transformation des segments en unités lexicales, soit à travers la consultation de lexiques, soit à l'aide de méthodes de correction. Nous avons vu dans la section 1.3.3 que pour certains contextes d'application, la correction de SMS notamment, cette phase de traitement

[5] Dans notre contexte de TALN, nous désignons comme expert, le linguiste-informaticien en charge de la mise en place et du paramétrage du système d'analyse.

conduisait à la génération d'un nombre important d'hypothèses concurrentes et de multiples erreurs d'interprétation qui affectaient ensuite l'efficacité des modules de traitement suivants.

Pour être en mesure de prioriser certaines unités lexicales ou de filtrer celles pouvant correspondre à des erreurs d'interprétation, différentes stratégies ont été envisagées pour générer les "éléments de décision" nécessaires à la mise en place d'une action de contrôle :

- fréquence d'apparition de l'unité lexicale ;
  Lorsqu'un corpus représentatif du domaine d'analyse est disponible, la fréquence d'apparition de chaque mot peut être calculée et ajoutée dans le lexique en tant que trait complétant la description des unités lexicales. Par exemple, toujours en se référant à la stratégie de correction de SMS, l'interprétation d'un segment tel que "pars" dans le message "2min pars ke G 1 match" proposera au moins 2 entrée du lexique "parse" et "parce", la première étant incorrecte. Dans un corpus SMS il est évidemment plus probable d'observer la forme "parce" en tant que partie de la locution "parce que" plutôt que la forme "parse" en tant qu'adjectif qualifiant une appartenance au peuple asiatique des Parsis.
- préférence syntaxique ;
  Lorsque plusieurs catégories morpho-syntaxiques associées à une unité lexicale sont en concurrence, le module d'analyse lexicale peut consulter un fichier de préférences, qui associent de manière heuristique une importance plus forte à certaines catégories par rapport à d'autres. Associé à une analyse de fréquence catégorielle, ce système de préférence permet de privilégier les catégories les plus fréquentes.
- compatibilité sémantique avec son voisinage.
  Les unités qui composent les lexiques utilisés par `TiLT` peuvent être reliées par l'intermédiaire d'une clé à une entrée d'un thesaurus, regroupant différentes descriptions et liens sémantiques. Une méthode spécifique a été développée pour calculer la distance sémantique d'un mot, ou plutôt des différents sens qui lui sont associés, avec ceux des mots avoisinants. Par l'intermédiaire d'un score représentant cette distance, il est ensuite possible de prioriser les unités lexicales les plus proches thématiquement de leur contexte local.

Contrôle syntaxique

La phase d'analyse syntaxique est communément considérée comme une étape du processus de traitement à la source d'une explosion combinatoire du nombre d'hypothèses générées et également d'erreurs d'interprétation. Devant la difficulté que constitue la spécialisation des règles de grammaire, des "éléments de décision" ont été intégrés pour évaluer et qualifier la pertinence des arbres syntaxiques de dépendance générés. Parmi ces informations distinctives ajoutées aux structures syntaxiques, nous avons :

- les cadres de sous-catégorisation ;
  Les informations de sous-catégorisation associées aux unités lexicales permettent d'identifier la structure des phrases et constituent donc des connaissances très importantes pour l'analyse syntaxique [Briscoe and Carroll, 1993]. Comme pour toute ressource linguistique, il est impossible de prétendre disposer de l'exhaustivité des cadres de sous-catégorisation de l'ensemble des unités lexicales rencontrées. Ainsi, pour des raisons de robustesse, ces connaissances ne sont pas exploitées en tant que contraintes mais en tant "qu'éléments de décision". L'information distinctive apportée par les sous-catégorisations se matérialise par un score associé à chaque arbre de dépendance qui représente le nombre de cadres de sous-catégorisation remplis par sa structure syntaxique.
- les contraintes sémantiques structurelles ;

La levée de nombreuses indéterminations syntaxiques repose sur des connaissances sémantiques. Lorsque que ces informations sont disponibles, les structures prédicatives des verbes permettent de contraindre certains attachements en fonction des concepts associés aux unités lexicales. L'information binaire apportée par cet "élément de décision", conformité ou non vis à vis des contraintes sémantiques, permet principalement de filtrer les unités lexicales non compatibles avec ces contraintes syntactico-sémantiques. "L'avocat mange l'avocat." est l'exemple jouet illustrant l'apport de ces connaissances, où la structure prédicative associée au verbe "manger" permet de ne sélectionner que les terminaux en position d'objet direct qui ont un concept relié à celui d'aliment.

– la forme de l'arbre, la taille de l'arbre et la reconnaissance des locutions ;
  Les hypothèses syntaxiques sont également évaluables selon différentes heuristiques comme la forme de l'arbre (attachement droit) ou la taille de l'arbre (attachement minimal), cette dernière propriété permet notamment de privilégier les arbres contenant des locutions, c'est-à-dire reposant sur un minimum de relations de dépendance.
– la conformité vis-à-vis d'un meilleur chemin probabiliste.
  Un algorithme d'étiquetage probabiliste par bigramme permet, lorsqu'un apprentissage sur un corpus représentatif a pu être effectué, d'identifier dans le treillis des hypothèses lexicales la succession de catégories morpho-syntaxiques la plus probable. Cette information syntaxique de surface est exploitée pour marquer à l'aide d'un "élément de décision" binaire les arbres respectant cette meilleure succession de catégorie morpho-syntaxiques définie à l'aide de connaissances probabilistes.

Le contrôle des arbres syntaxiques en dépendance, qu'il s'agisse de sélection d'une meilleure hypothèse ou de les classer, repose donc sur l'exploitation de ces informations distinctives. Nous verrons dans la section suivante que ces connaissances ne sont cependant pas exploitées de manière optimale par les stratégies de contrôle envisagées jusqu'ici, mais qu'il en est également de même pour le cas particulier de la chaîne de traitement `TiLT`.

À travers ces deux exemples, on constate donc que pour pallier le manque de caractère distinctif entre les hypothèses concurrentes, des "éléments de décision" ont été ajoutés pour évaluer les objets linguistiques construits par un module de traitement. De différentes natures : statistique (probabilité, fréquence), numérique (score, coût) ou binaire (respect d'une propriété), ces informations résultent de l'usage de ressources supplémentaires ou simplement d'heuristiques visant par exemple à considérer un trait linguistique (par ex. : la catégorie morpho-syntaxique) comme un "élément de décision".

Le contrôle des hypothèses concurrentes étant un phénomène fortement récurrent et affectant la plupart des modules de traitement, nous aurions pu illustrer l'intégration "d'éléments de décision" sur d'autres exemples. Nous verrons lors du chapitre 5 qu'un processus d'extraction de couples candidats `antécédent / reprise anaphorique` a été développé en se basant sur les informations produites par `TiLT`, et que la séparation des couples corrects et incorrects repose sur l'intégration "d'éléments de décision" hétérogènes de par leur nature (statistique, symbolique, numérique) et leur type (lexical, syntaxique, sémantique, pragmatique).

2.2.4 Des "éléments de décision" aux critères de comparaison

À travers cette section, nous avons vu que l'évaluation d'hypothèses concurrentes à un moment donné du processus de traitement repose sur la prise en compte de connaissances supplémentaires. Nous avons vu que sous le terme générique "d'élément de décision" emprunté à Gérard Sabah, nous englobons l'ensemble des informations distinctives utilisées pour la diffé-

rentiation et l'évaluation des hypothèses concurrentes. Décrits ou non comme des stratégies de contrôle, de nombreux travaux reposent sur l'interprétation de ces informations distinctives en vue d'évaluer la pertinence des différentes hypothèses.

Matérialisée par une valeur numérique (score, probabilité) ou binaire (conforme / non conforme), chaque information distinctive apporte un point de vue particulier (lexical, syntaxique, sémantique, pragmatique, empirique, statistique, etc.) pour juger de la pertinence relative des hypothèses. Afin de poursuivre dans le paradigme décisionnel et afin de rapprocher les éléments manipulés de la terminologie établie dans ce domaine, on constate que les différentes informations distinctives utilisées pour une tâche de contrôle peuvent être envisagées comme des critères de comparaison. En se référant au TLFI[6], un critère correspond à un "caractère, principe, élément auquel on se réfère pour juger, apprécier, définir quelque chose (...)". La définition donnée par Bernard Roy et Denis Bouyssou dans un ouvrage spécialisé sur les méthodes de prise de décision [Roy and Bouyssou, 1993] apporte un complément sur la nature de ces informations : "Pour l'essentiel, un critère vise à résumer, à l'aide d'une fonction, les évaluations d'une action sur diverses dimensions pouvant se rattacher à un même axe de signification, ce dernier étant la traduction opérationnelle d'un point de vue, au sens usuel du terme".

Comme pour la plupart des problèmes décisionnels, on constate que l'évaluation des hypothèses concurrentes repose rarement sur la prise en compte d'un seul axe de signification. Nous avons en effet vu à travers deux exemples que pour chaque cas d'indétermination rencontré au cours du processus d'analyse, plusieurs types de critères avaient été envisagés, chacun d'eux correspondant à un axe de signification particulier que nous avons jusqu'ici désigné comme le type du critère (lexical, syntaxique, sémantique, pragmatique).

Ces critères de comparaison permettant d'évaluer la pertinence relative des hypothèses concurrentes constituent la première réponse à la question soulevée par la définition d'une stratégie de contrôle : "Quelles sont les connaissances de contrôle ?" (section 2.1.1).

Ainsi, les critères de comparaison et plus précisément les performances de critères associées aux hypothèses concurrentes sont une représentation des connaissances nécessaires à la résolution d'un problème de contrôle. La nature et le type de ces critères sont évidemment étroitement liés au problème concerné [Bachimont, 1992] (page 31), il s'agit bien de représentation de connaissances exploitable pour le contrôle de certaines hypothèses et ces critères ont notamment un sens parce qu'ils sont exploités dans un contexte d'analyse précis.

Nous allons voir dans la section suivante que l'efficacité des stratégies de contrôle déployées repose sur une exploitation optimale de l'information distinctive apportée par chacun des critères de comparaison disponibles lors d'une étape de contrôle.

2.3 Une meilleure exploitation des critères de comparaison

Lors des deux premières sections de ce chapitre, nous avons proposé de modéliser les stratégies de contrôle en tant que processus décisionnels. La finalité de ce processus, c'est-à-dire la prise de décision d'une action de contrôle (tri, classement, sélection), repose sur une évaluation des hypothèses concurrentes. Dans la section précédente 2.2, nous avons vu que la différentiation des hypothèses, qui apparaissent intrinsèquement comme concurrentes, s'appuie sur l'intégration de

[6] Trésor de la Langue Française Informatisé http://atilf.atilf.fr

connaissances supplémentaires lors du processus d'analyse, connaissances que nous considérons comme des critères de comparaison.

Le regroupement en classes d'équivalence ou la construction d'un ordre sur l'ensemble des hypothèses concurrentes se base sur l'interprétation de ces critères de comparaison. Cette section vise à montrer que les méthodes envisagées en ce sens ne permettent pas d'exploiter pleinement l'information portée par les différents critères de comparaison disponibles lors d'une étape de contrôle. Ces constats nous permettront d'introduire la méthodologie de contrôle que nous proposons, méthodologie qui cherchera a obtenir de plus une certaine généricité, c'est-à-dire l'indépendance de la méthode vis-à-vis des particularités des différents "points de décision" et de la problématique de contrôle visée.

2.3.1 Limite des approches monocritère de contrôle

Bien que la nécessité d'évaluer et de contrôler à l'aide de méthodes dédiées le processus de génération d'interprétations linguistiques semble communément acceptée, les stratégies de contrôle envisagées restent fortement spécialisées dans la résolution d'un cas précis d'indétermination. Le contrôle est alors souvent relégué en tant que "rustine" ou technique d'optimisation d'une méthode d'analyse, et malgré le caractère récurrent et omniprésent de cette problématique, il est très délicat de dégager des travaux existants un ensemble de méthodologies concurrentes et de les évaluer.

Ainsi, les techniques de contrôle, et notamment celles décrites dans la section 2.2, sont systématiquement conçues et présentées soit comme des extensions de méthodes particulières de traitement : [Charniak, 2005] comme une amélioration de l'analyseur syntaxique de [Charniak, 2000], [Blache and Rauzy, 2006] comme un contrôle d'une étape précise d'application des grammaires de propriétés de [Blache, 2000] ; soit comme des techniques étroitement liées à une tâche : désambiguïsation du sens des mots, sélection d'une meilleure traduction, correction et réaccentuation [Yarowsky, 1994], etc.

De plus, nous avons vu dans la section 2.1 que l'exploitation des critères de comparaison par les méthodes de contrôle pouvait se matérialiser par trois actions différentes : le tri, le classement ou la sélection. L'information apportée par un critère de comparaison est indépendante de l'usage qui en est fait et de l'action de contrôle visée. Cependant, chaque approche développée pour exploiter ces critères ne concernait qu'une seule des trois actions de contrôle envisageables et aucune de ces approches ne permet de tester l'efficacité d'actions différentes pour un cas précis d'indétermination. Il serait par exemple intéressant de déterminer si le contrôle est suffisamment fiable pour ne conserver qu'une meilleure hypothèse ou si N hypothèses doivent être propagées.

La spécialisation des différentes stratégies de contrôle connues s'explique également par le fait que la plupart des travaux conduits par des équipes de recherche en TALN est focalisée sur une tâche particulière d'analyse. À ce jour, le nombre de systèmes complets de TALN capables d'effectuer une analyse selon différents niveaux de traitement (voir chapitre 1) est très faible. Le besoin de disposer d'une méthodologie de contrôle réutilisable et applicable lors des différents "points d'embarras" pouvant subvenir lors du processus d'analyse ne s'est pas fait ressentir et donc aucune démarche de généralisation n'a été entreprise.

Outre ces limites méthodologiques, on constate que malgré la diversité des critères de comparaison envisageables pour un unique cas de contrôle, les approches proposées ont longtemps été focalisées sur l'exploitation d'un seul critère de comparaison. À titre d'exemple, de longs débats ont été animés par différents chercheurs en sciences cognitives pour déterminer quelle

propriété syntaxique devait être respectée pour choisir comment attacher au reste de l'analyse les syntagmes découverts au cours de la lecture. Qu'il s'agisse d'attachement droit, d'attachement minimal, de rôle sémantique ou d'anticipation, chacun des protagonistes de ces débats a mis en avant l'apport d'une propriété avant qu'un autre la contredise avec une série de contre-exemples [Schubert, 1984] [Wilks, 1985]. De même, les approches de traitement statistique ont principalement exploité la fréquence d'observation des structures linguistiques reconnues comme seul critère de comparaison pour le classement des hypothèses générées. La tâche de contrôle plus communément définie sous le terme de désambiguïsation du sens des mots a également été longtemps basée sur la prise en compte d'une unique heuristique [McCarthy *et al.*, 2004], méthode [Agirre and Rigau, 1996] ou source de connaissance [Yarowsky, 1992] (thesaurus ou dictionnaire).

Cependant, l'efficacité d'une stratégie de contrôle basée sur un unique critère de comparaison se limite aux cas couverts par cette information distinctive. Comme l'ont montré de nombreux travaux en sciences cognitives [Gibson and Pearlmutter, 1998] et comme nous le verrons plus concrètement dans le chapitre 5, les choix émis au cours d'un processus d'analyse reposent sur plusieurs sources de connaissances, plusieurs points de vue et donc plusieurs critères de comparaison. Ne considérer qu'un seul point de vue de jugement pour statuer de la pertinence des hypothèses concurrentes constitue donc la limite prépondérante de bon nombre de méthodes de contrôle.

2.3.2 Limites des approches multicritère de contrôle

Constatant le manque de robustesse et d'efficacité des approches de contrôle monocritère, des méthodes basées sur la combinaison de sources de connaissances et donc de plusieurs critères de comparaison ont été développées. Quelle que soit la nature du problème décisionnel rencontré, la prise en compte de plusieurs points de vue permet souvent d'atteindre une meilleure compréhension des alternatives à évaluer et d'arriver à une prise de décision finale plus juste. En effet, nous avons signalé précédemment que pour un "point de décision" précis, il ne semblait pas exister de critère suffisamment discriminant pour guider à lui seul la prise de décision. L'information portée par un critère de comparaison doit donc être considérée comme une source d'évaluation des hypothèses pouvant être erronée ou incomplète. Ces erreurs de jugement proviennent notamment des différents contre-exemples que l'on peut trouver pour chaque propriété énoncée. De plus, les critères basés sur des ressources linguistiques complémentaires peuvent également être dans l'incapacité de porter un jugement si les données nécessaires ne sont pas disponibles. Quant aux critères statistiques, ils reprennent également les défauts liés aux approches basées sur corpus. Ils apparaissent en effet comme des sources de jugement très efficaces pour identifier les structures linguistiques récurrentes, mais s'avèrent souvent inefficaces pour traiter des phénomènes minoritaires ou exceptionnels.

La prise en compte de plusieurs critères de comparaison pour une tâche de contrôle vise à rattraper l'absence ou les erreurs de jugement commises individuellement par certains critères. Ainsi, disposer d'un ensemble cohérent de critères de comparaison permet à la fois d'obtenir un jugement plus robuste et fiable, notamment à travers la confirmation d'un jugement par plusieurs critères, mais également de récupérer des erreurs individuelles de jugement en recherchant un compromis entre les informations contradictoires.

Dans de nombreux contextes de contrôle, des travaux ont montré l'apport de la prise en compte de plusieurs critères de comparaison à la fois en terme de justesse, mais également en

terme de compréhension du problème et d'interprétation des décisions émises. La désambiguïsation du sens des mots constitue le cas de contrôle où la nécessité de prendre en compte plusieurs critères de comparaisons a été largement démontrée [Audibert, 2003], [Wilks, 1998], [McRoy, 1992], [Dang and Palmer, 2002], [Rigau *et al.*, 1997] ou encore [Rosso *et al.*, 2003]. Comme le montrent ces travaux, décider du sens le plus pertinent à affecter à une unité lexicale repose sur des connaissances de nature variée : lexicale, syntaxique, sémantique et pragmatique. Cet apport des stratégies multicritère a également été souligné pour d'autres tâches de contrôle telles que l'analyse syntaxique de surface [Ratnaparkhi, 1996], l'analyse syntaxique profonde [Charniak, 2005] ou encore l'extraction d'informations pour caractériser un document [Chieu and Ng, 2002].

Pour que la combinaison de plusieurs critères de comparaison apporte réellement une évaluation plus fiable et plus robuste des hypothèses et non une complexification désuète du processus de contrôle, il est nécessaire à la fois de disposer de critères de comparaison pertinents [Audibert, 2007] mais également d'une méthode d'agrégation efficace. Une méthode "d'agrégation multicritère consiste à synthétiser des informations traduisant des aspects ou des points de vues différents et parfois conflictuels au sujet d'un même ensemble d'objets" [Grabisch and Perny, 2002]. Elle doit donc à la fois respecter l'information apportée individuellement par chaque critère mais également fournir une évaluation finale résultant d'un compromis entre les critères divergents. Si comme le souligne [Audibert, 2007], la pertinence des critères choisis conditionne fortement la qualité du résultat de l'agrégation, la façon dont chaque information est interprétée joue également un rôle important lors de la comparaison des hypothèses concurrentes. En fonction des particularités du contexte de contrôle, les critères utilisés peuvent en effet apparaître d'importance variable. Certains prévalant par rapport à d'autres, il faut que la méthode d'agrégation puisse intégrer cette importance relative. De plus, les critères à agréger sont souvent hétérogènes, de par leur nature mais également leur type (binaire, numérique, statistique) et le domaine de définition de l'information distinctive qu'ils apportent. Ainsi, comparer une hypothèse (problématique de classement ou de sélection) avec une autre ou avec le profil d'une classe (problématique de tri) à partir de leurs performances respectives sur un critère binaire ou sur un critère numérique ne s'effectuera pas de la même façon. Par exemple, considérons la comparaison de deux arbres syntaxiques à l'aide de deux critères, l'un associant à chaque arbre un score numérique désignant la fréquence moyenne d'observation des attachements qui le compose et l'autre calculant le nombre de cadres de sous-catégorisation remplis par la construction. Il apparaît évident qu'une différence même faible entre les deux hypothèses sur le critère de sous-catégorisation est fortement significative alors qu'une différence très faible sur la fréquence d'observation ne l'est pas forcément. Ces paramètres, que nous désignons en tant que préférences, reposent sur des connaissances qu'un expert du domaine peut formuler à l'aide de ses intuitions ou de son expérience et visent à obtenir une meilleure exploitation de l'information apportée par chacun des critères.

On constate cependant que les méthodes d'agrégation multicritère exploitées jusqu'ici, principalement issues de l'apprentissage automatique (machines à vecteurs supports (SVM), réseaux bayésiens, liste de décisions, entropie maximale, etc.), ne permettent pas d'intégrer facilement ce genre de connaissance. Au contraire, le processus d'agrégation et donc de décision est très souvent difficilement interprétable et les erreurs de jugement deviennent difficiles à tracer, ce qui est notamment le cas pour les méthodes à base de SVM ou de réseaux bayésiens.

Les autres méthodes, comme celles basées sur le principe d'entropie maximale, sont principalement focalisées sur les aspects liés à l'exploitation du corpus d'apprentissage en vue de

quantifier les différents traits considérés comme des critères de comparaison, ceci au détriment du processus d'agrégation de ces informations. Les stratégies d'agrégation comme celles utilisées dans [Charniak, 2005] s'appuient sur des méthodes simples telles que les sommes pondérées, qui outre le fait qu'elles exigent de disposer de critères commensurables, se basent sur une hypothèse forte de complémentarité et de compensation des critères qui apparaît comme insatisfaisante dans de nombreux cas. En effet, par addition ou multiplication des performances obtenues par une hypothèse sur les différents critères exploités, on considère qu'une évaluation très négative sur un critère peut être compensée par des bonnes performances sur les autres critères. De plus, ces méthodes (entropie maximale, SVM, réseaux bayésiens, etc.) reprennent également les limites énoncées précédemment sur l'usage de critères uniquement statistiques, à savoir la dépendance de ces stratégies vis à vis de la disponibilité de corpus d'apprentissage exploitables.

D'autres méthodes plus simples, telles que les listes de décision [Yarowsky, 1994] ou les ordres lexicographiques [Nastase and Szpakowitcz, 2001], conduisent également à une faible exploitation de la complémentarité des différents critères de comparaison exploités. La décision finale émise à l'aide de ce genre d'approche repose fréquemment sur le jugement apporté par le critère défini comme le plus important, qui devient alors un critère dit dictateur. En effet, il suffit à une hypothèse d'être sensiblement préférée selon un critère pour être privilégiée, même si elle apparaît nettement moins performante sur l'ensemble des autres critères disponibles.

De plus, on constate que les critiques émises sur les approches monocritères lors de la section 2.3.1, notamment sur leur spécialisation, s'appliquent également à ces approches multicritères. En effet, bien que des méthodes telles que les SVM, les réseaux bayésiens ou l'entropie maximale puissent être appliquées sur divers cas de contrôle, elles ne répondent qu'à une seule des trois problématiques de contrôle présentées lors de la section 2.1. Sans remettre en cause leur efficacité, nous verrons dans la section suivante que nos objectifs, principalement focalisés sur la recherche d'une méthodologie générique et flexible de contrôle, rendent ces méthodes peu envisageables. Nous verrons également que du point de vue de l'agrégation de critères hétérogènes, les méthodes envisagées jusqu'ici ont le principal défaut de ne pas prendre en compte des connaissances *a priori* sur le domaine formulables par un expert.

Lors de la définition de la notion de contrôle section 2.1.1, nous avons effectué un rapprochement avec des travaux plus génériques sur le contrôle des systèmes à base de connaissances et donc sur le contrôle de systèmes d'IA [Bachimont, 1992]. Ce contrôle en IA a, d'après nos connaissances, uniquement été relié aux architectures de tableaux noirs. Le contrôle de la résolution de problèmes de reconnaissance vocale par le système `Hearsay II` [Erman *et al.*, 1980] est une illustration de ces travaux sur le contrôle en IA. L'approche de contrôle des objets concurrents construits et déposés sur le tableau noir d'`Hearsay` a très tôt mis en évidence l'intérêt d'une approche multicritère. Les concepteurs de cette approche et des modules qui génèrent les hypothèses potentiellement concurrentes ont proposé d'adopter cinq points de vue de jugement complémentaires pour déterminer la pertinence de ces différentes hypothèses. Ainsi les modules de traitement étaient en charge de déterminer la pertinence de chaque hypothèse construite sous des points de vue de compétitivité, de crédibilité, d'importance, d'efficacité et de satisfaction du but. Cette stratégie de contrôle purement fonctionnelle s'est avérée très délicate à mettre en œuvre et à maintenir.
C'est pourquoi, comme nous allons le constater par la suite, l'approche de contrôle que nous avons construite s'appuie sur une stratégie la plus déclarative possible pour manipuler les connaissances de contrôle disponibles.

2.3.3 Décision ou aide à la décision : vers plus de généricité et de flexibilité

Dans la première section de ce chapitre, nous avons montré que les stratégies visant à réduire la proportion d'hypothèses erronées et à éviter leur propagation au cours du processus d'analyse pouvaient être généralisées sous la notion de contrôle. Appuyée sur des travaux plus anciens, nous avons vu que la démarche de contrôle correspondait à un processus décisionnel ayant pour objectif d'établir un tri, un classement ou une sélection sur l'espace des hypothèses concurrentes à partir d'une évaluation de leur pertinence respective. Lors de la seconde section, nous avons vu que cette évaluation reposait sur l'intégration de critères de comparaison et que l'obtention d'une décision de contrôle fiable et robuste nécessitait la prise en compte de plusieurs critères complémentaires. Les approches existantes de contrôle basées sur l'intégration de connaissances spécifiques de contrôle ne constituent cependant que des propositions de stratégies très spécifiques à une tâche et une problématique données. De plus, ces méthodes ne permettent pas d'intégrer des connaissances expertes décrivant la façon dont les critères doivent être exploités, ce qui conduit souvent à la faible exploitation de la complémentarité des critères disponibles.

Nos travaux s'articulent autour de deux axes, l'un théorique et méthodologique, l'autre pratique et opérationnel. Ainsi, après avoir rapproché le contrôle de la théorie de la décision, nous nous intéressons désormais à la mise en place d'une méthodologie de contrôle et d'outils opérationnels. Contrairement aux méthodes de contrôle envisagées jusqu'à ce jour, nos travaux se concentrent principalement sur la recherche de `généricité`. Cette propriété se matérialise par l'obtention d'une méthode d'agrégation des critères de comparaison qui soit applicable sur les différents cas d'indétermination pouvant apparaître lors d'un processus d'analyse, mais qui puisse également être utilisable lorsqu'aucun corpus d'apprentissage n'est disponible. Ce qui signifie que cette méthode doit être suffisamment `paramétrable` et `flexible` pour qu'une stratégie de contrôle puisse être établie à partir de connaissances du domaine formulées par un expert. Comme nous le verrons dans le chapitre consacré aux expérimentations et évaluations (Chap. 5), l'usage de stratégies de contrôle vise principalement à réduire la surgénération d'hypothèses. Ces stratégies doivent également être en mesure d'apporter des éléments importants de compréhension du processus de génération des hypothèses, notamment sur les possibilités de factoriser certaines structures linguistiques à l'origine de sur-génération ou au contraire de spécialiser et contraindre certaines règles. Il apparaît donc important que les évaluations et les décisions émises soient `traçables`, c'est-à-dire interprétables et compréhensibles *a* posteriori.

Nous verrons ensuite que la recherche de généricité pour une méthode de contrôle prend tout son sens lorsque l'on cherche à l'intégrer dans une chaîne de traitement paramétrable comme `TiLT`. La méthode envisagée a donc donné lieu à l'implémentation d'un module dédié au contrôle du processus classique d'analyse conduit par `TiLT`. Afin d'attester de la généricité de cette approche, nous l'appliquerons à deux cas concrets d'expérimentation. Les interprétations des résultats obtenus sur ces cas d'indétermination témoigneront de la traçabilité du processus décisionnel de contrôle. Quant à la flexibilité et à la paramétrabilité de la méthode, nous verrons que la méthode proposée laisse une place prépondérante à un expert du domaine, lui permettant d'introduire ses connaissances et intuitions pour la mise en place d'une stratégie de contrôle. Cependant, cette méthode doit également être capable d'exploiter pleinement un corpus représentatif du cas de contrôle concerné, lorsque celui-ci est disponible.

Ces objectifs issus du constat des limites des approches existantes nous entraînent à considérer la méthodologie de contrôle non pas comme un procédure décisionnelle automatique, mais

comme un ensemble d'outils guidant des experts vers la mise en place de stratégies de contrôle adaptées aux particularités d'un cas de "point de décision" et exploitant de manière optimale les critères disponibles. Cette démarche nous a ainsi conduit à rapprocher deux domaines de recherche jusqu'ici sans intersection, le TALN à travers sa problématique de contrôle et l'Aide Multicritère à la Décision (AMCD) fournissant la méthodologie souhaitée.

En informatique plus généralement, ces besoins méthodologiques de soutien des experts lors de la mise en place de stratégies décisionnelles sont désignés sous le terme d'aide à la décision. Nous allons ainsi constater que les aspects méthodologiques et architecturaux de nos travaux comportent des similitudes avec ce domaine [Chaudhuri and Dayal, 1997], notamment pour des questions :

- de stockage des éléments décisionnels ;
- d'accès aux connaissances ;
- de traitement des connaissances décisionnelles ;
- d'interprétation des suggestions de décision.

3

L'aide multicritère à la décision et le TALN

Sommaire

3.1 L'Aide MultiCritère à la Décision (AMCD) 44
3.1.1 L'AMCD en tant qu'extension pragmatique de la théorie de la décision 44
3.1.2 Formalisation d'un problème d'AMCD 45
3.1.3 Les approches en AMCD 46
3.2 Les approches par surclassement 49
3.2.1 Modélisation des préférences et connaissances expertes 49
3.2.2 Le surclassement comme relation générique de comparaison 51
3.2.3 `ELECTRE III` et `ELECTRE TRI` 53
3.3 Une approche par surclassement pour le contrôle des indéterminations 62
3.3.1 Particularités du contexte décisionnel 62
3.3.2 Une méthodologique axée sur la généricité 63
3.3.3 Adaptation et simplification d'aspects techniques 63

La recherche d'une approche de contrôle générique et flexible nous a conduit à envisager l'usage de méthodes spécialisées dans la prise de décision à partir de critères multiples hétérogènes. Ce chapitre introduit dans une première section le domaine de l'Aide MultiCritère à la décision (AMCD) et nous verrons dans une deuxième section que les méthodes par surclassement apportent les notions et les propriétés les plus prometteuses pour remplir l'ensemble des propriétés attendues (section 2.3.3). Lors de la troisième et dernière section de ce chapitre, nous verrons comment l'approche d'AMCD par surclassement s'applique à notre problématique de contrôle et quelles sont les particularités de notre contexte décisionnnel.

3.1 L'Aide MultiCritère à la Décision (AMCD)

3.1.1 L'AMCD en tant qu'extension pragmatique de la théorie de la décision

Nous avons vu précédemment que la prise de décision, dans notre cas la mise en place d'une action de contrôle (tri, sélection, classement), n'était que le résultat d'un processus rationnel plus complet, notamment orienté autour de l'évaluation des différentes hypothèses en concurrence. Principalement à des fins économiques ou stratégiques (politique et militaire), des travaux ont cherché à étudier et décrire ces processus conduisant à la prise de décision [Hansson, 1994]. Depuis les premiers travaux théoriques apparus à la fin du XVIII^{eme} siècle attribués à Condorcet, l'étude des processus décisionnels revêt un caractère plus pragmatique et méthodologique. Ces études se sont intéressées aux différentes notions sous-jacentes à notre faculté décisionnelle, conduisant ainsi à différents domaines de recherche spécialisés tels que la recherche opérationnelle, la modélisation des préférences, la théorie du choix social, la théorie du vote, etc.

Située à l'intersection de ces différents domaines de recherche, l'AMCD s'est principalement développée dans la seconde partie du XX^{eme} siècle avec comme ambition d'étudier et de modéliser les processus décisionnels basés sur plusieurs sources de jugement. L'objectif de ces travaux était d'aboutir à un ensemble de méthodes privilégiant les aspects pragmatiques plutôt que théoriques. Ce domaine de recherche à part entière "vise, comme son nom l'indique, à fournir à un décideur des outils lui permettant de progresser dans la résolution d'un problème de décision où plusieurs points de vue, souvent contradictoires, doivent être pris en compte." [Vincke, 1989]. Cette définition illustre parfaitement l'esprit et l'objectif de l'AMCD, où les méthodes développées ne cherchent pas à tout prix à obtenir une solution "optimale" à un problème décisionnel, mais bien à guider un décideur vers une exploitation efficace des différents critères envisagés pour conduire à une prise de décision la plus conforme possible vis à vis des connaissances disponibles. Cette recherche de compromis plutôt que d'optimalité ainsi que cette démarche orientée autour du décideur font de ces méthodes des approches d'aide plutôt que de résolution automatique.

Bien que ces méthodes soient souvent qualifiées de peu axiomatiques ou de mathématiquement peu rigoureuses, leur efficacité ainsi que la pertinence des résultats générés ont été illustrées dans divers contextes applicatifs (voir [Vincke, 1989, p.154-155]) tels que l'économie, l'environnement, l'écologie, les transports, les sciences et plus en rapport avec notre domaine de recherche : le classement des résultats d'un moteur d'indexation [Farah and Vanderpooten, 2006].

L'efficacité des méthodes d'AMCD pour traiter un problème décisionnel basé sur plusieurs points de vue de jugement provient principalement de l'importance accordée aux acteurs décisionnels. En effet, en tant que procédures d'aide, les différentes méthodes développées constituent une interface entre le décideur et la prise de décision visant à représenter finement les composants décisionnels (section 3.1.2) et les connaissances (section 3.2.1) dont dispose le décideur sur

le problème.

Initialement développée en vue de constituer des outils à la disposition d'acteurs décisionnels humains : le décideur et l'homme d'étude, l'AMCD attribue une place prépondérante aux experts du domaine dans la formalisation du contexte décisionnel. Ainsi, l'AMCD, à travers les différentes méthodes qu'elle propose, apparaît comme un axe de recherche prometteur pour la mise en place d'une stratégie de contrôle générique. En effet, en offrant à un expert du domaine (linguiste ou informaticien) la possibilité d'introduire ses connaissances sur le contexte de contrôle dans le processus décisionnel, la méthode peut être adaptée et spécialisée en exploitant ces connaissances *a priori* sans dépendre d'un corpus d'apprentissage. Ainsi, notre démarche expérimentale visant à rapprocher l'AMCD et le TALN n'a pas comme premier objectif l'obtention de résultats probants sur la résolution d'un cas d'indétermination particulier, mais bien de valider l'intérêt des méthodes d'AMCD pour la mise en place d'une stratégie générique de contrôle. Nous verrons également dans quelle mesure ce choix est pertinent pour le système particulier de traitement `TiLT`.

3.1.2 Formalisation d'un problème d'AMCD

Avant de procéder à une description de ses composantes, nous proposons de citer la définition d'un problème multicritère donnée par Philippe Vincke dans [Vincke, 1989, p. 54][7] : "Un problème de décision multicritère est une situation où, ayant défini un ensemble H d'hypothèses et une famille G cohérente de critères sur H, on désire

- soit déterminer un sous-ensemble d'hypothèses considérées comme les meilleures vis-à-vis de G (problème de choix)
- soit partitionner H en sous-ensembles suivant des normes préétablies (problème de tri)
- soit ranger les actions de H de la meilleure à la moins bonne (problème de rangement)."

Quelle que soit la problématique à résoudre, un cas d'application d'une méthode d'AMCD repose sur deux composants élémentaires : l'ensemble des hypothèses et l'ensemble des critères.

Les actions/hypothèses

Souvent désigné sous la notion d'action, de solution ou d'alternative dans les ouvrages d'AMCD, nous employons le terme d'hypothèse pour désigner les éléments d'un ensemble d'objets concurrents de même nature qui doivent être évalués lors du processus de décision. Bien que des méthodes aient été développées pour les cas où cet ensemble ne peut être défini qu'en compréhension, nous nous intéresserons uniquement aux ensembles d'hypothèses $H : \{h_1, h_2, ..., h_n\}$ finis.

Bien qu'il soit parfois difficile à établir pour certains contextes décisionnels, l'ensemble H est établi comme une donnée initiale de notre problème et est composé des hypothèses concurrentes générées par un module de traitement.

[7] Afin d'harmoniser l'ensemble des notations et notions utilisées dans ce document, nous avons quelques peu modifié la définition originale en transformant "un ensemble A d'actions" et "une famille F cohérente de critères" par respectivement, "un ensemble H d'hypothèses" et "une famille G cohérente de critères"

Critères et familles de critères

Notre problème du contrôle se rapporte à un problème d'évaluation dans lequel une comparaison d'hypothèses doit être effectuée. Nous avons vu que les différentes hypothèses de l'ensemble H n'étaient pas différentiables intrinsèquement et que la mise en place d'une stratégie de contrôle reposait sur l'intégration d'informations distinctives supplémentaires. Afin de se rapprocher du paradigme de l'AMCD, nous avons qualifié chaque information distinctive de critère de comparaison. La définition donnée par `Bernard Roy` dans [Roy and Bouyssou, 1993] (page 46) de la notion de critère décrit parfaitement ces connaissances supplémentaires ajoutées au cours du processus d'analyse : "Pour l'essentiel, un critère vise à résumer, à l'aide d'une fonction, les évaluations d'une hypothèse sur diverses dimensions pouvant se rattacher à un même "axe de signification", ce dernier étant la traduction opérationnelle d'un "point de vue" au sens usuel du terme."

Cette définition introduit un aspect formel intéressant de la notion de critère, à savoir son aspect fonctionnel. Plus formellement, un critère k peut donc être représenté par une fonction g_k associant à une hypothèse h_i une valeur notée $g_k(h_i)$ dans son domaine de définition. Cette valeur que nous appelons performance, permet d'établir une comparaison entre deux hypothèses h_i et h_j. Ainsi, si $g_k(h_i) > g_k(h_j)$ alors on peut qualifier l'hypothèse $g_k(h_i)$ de préférable par rapport à $g_k(h_j)$ selon le critère k. Un critère doit donc correspond à une fonction continue croissante permettant de matérialiser par une valeur réelle l'évaluation d'une hypothèse selon un point de vue particulier de jugement.

Chaque point de vue n'apportant qu'une caractérisation et une différentiation partielle des hypothèses concurrentes, une comparaison complète repose sur la prise en compte d'un ensemble de critères $G : \{g_1, g_2, ..., g_m\}$ appelé souvent famille de critères. Il convient de constituer ce regroupement de critères de manière efficace afin d'obtenir ce que [Roy and Bouyssou, 1987] nomme une famille cohérente de critères. Pour résumer cette notion, il s'agit d'obtenir une représentation la plus complète possible des différents axes pouvant qualifier une hypothèse, tout en veillant à ce que chaque critère utilisé apporte une information supplémentaire et complémentaire par rapport aux autres sans redondance.

Chaque hypothèse de l'ensemble H est alors associée à un vecteur de valeurs réelles, où chaque valeur correspond au résultat de l'évaluation de l'hypothèse sur un critère. L'ensemble de ces valeurs est également nommé vecteur de performances. Le processus de décision, c'est à dire l'application d'une méthode d'agrégation multicritère, repose sur l'exploitation et la comparaison de ces vecteurs de performances.

3.1.3 Les approches en AMCD

Un problème d'aide multicritère à la décision vise donc à répondre à une des trois problématiques citées précédemment, ceci à partir d'une synthèse (agrégation) des critères $G : \{g_1, g_2, ..., g_m\}$ qualifiant les différentes hypothèses de l'ensemble H.
L'ensemble des méthodes d'AMCD sont fréquemment regroupées en trois approches :

- les approches par critère unique de synthèse ;
- les approches interactives ;
- et les approches par surclassement.

Critère unique de synthèse

Les méthodes appartenant à cette approche par critère unique de synthèse sont les plus connues. Également désignées sous le terme de théorie de l'utilité multiattribut par P. VINCKE dans [Vincke, 1989], ces méthodes construisent une évaluation synthétique de chaque hypothèse à partir de l'ensemble des critères qui la qualifie. La méthode la plus connue est sans doute la somme pondérée, utilisée par exemple pour synthétiser les différentes notes obtenues par un étudiant. Une telle méthode d'agrégation par somme pondérée associe à chaque hypothèse h_i de l'ensemble H un critère de synthèse $F(h_i)$ calculé de la façon suivante :

$$F(h_i) = \sum_{j \in G} w_j . g_j(h_i)$$

où w_j constitue un paramètre optionnel de pondération.

Qu'elles exploitent un modèle additif ou multiplicatif, ces méthodes reposent sur une fonction F : $F(g_1, g_2, ..., g_m)$ que l'on cherche à maximiser et qui à travers cette recherche d'optimum global agrège l'ensemble des points de vue disponibles.

La comparaison de deux hypothèses h_i et h_j est alors très simple, dans la mesure où il suffit d'établir un classement ($F(h_i) > F(h_j)$) ou un regroupement ($F(h_i) = F(h_j)$) sur la base de leur critère de synthèse.

Bien que faciles à mettre en œuvre, de par leur simplicité et leur efficacité, ces méthodes sont cependant inapdaptées dans de nombreux contextes décisionnels. En effet, ces approches par critère unique de synthèse nécessitent de disposer de *vrais-critères* dont les valeurs soient commensurables. Bien que les différences entre ces types de critères seront discutées au cours de la section 3.2.1, cette restriction aux *vrais-critères* signifie que la moindre différence de valeur sur un critère est significative, sans prise en compte d'imprécision et d'incertitude dans l'interprétation de ces valeurs. Si l'on revient aux exemples de cas de contrôle introduits lors de la section 1.3.3, on constate aisément qu'il sera délicat d'agréger par l'intermédiare de ces méthodes des critères que nous avons qualifiés de binaires (respect ou nom d'une propriété) et des critères numériques (statistiques, scores, etc.). En effet, la normalisation de ces données sur une échelle commune n'est pas toujours évidente.

Les approches interactives

Les méthodes suivant une approche dite interactive [Vanderpooten, 1989] mettent en étroite relation un ou des décideurs et un homme d'étude, terme désignant un spécialiste des processus décisionnels. Ces méthodes ne s'intéressent pas principalement aux aspects calculatoires mais davantage au protocole de communication qui peut se mettre en place entre les différents acteurs décisionnels. En effet, à travers une succession de phases expérimentales, l'objectif de ces méthodes est d'atteindre progressivement un système décisionnel conforme aux attentes du décideur. Ainsi, une succession de décisions locales sont prises, puis leurs résultats sont évalués afin de procéder à des raffinements du modèle décisionnel jusqu'à l'obtention de résultats acceptables.

Étant donné que nous cherchons à mettre en place des outils exploitables directement par des décideurs et que ces outils ne constitueront jamais des systèmes experts suffisament efficaces pour remplacer un homme d'étude, nous n'avons pas effectué une étude plus approfondie de ces méthodes.

Les approches par surclassement

L'approche par critère unique de synthèse s'appuie sur des hypothèses fortes (existence d'une fonction objectif, commensurabilité des critères, etc.) pour construire un classement des hypothèses. Si l'on raisonne en terme de relation entre les hypothèses, le classement obtenu intègre deux types de relations : la préférence et l'indifférence. La première désigne une situation de supériorité stricte d'une hypothèse par rapport à une autre ($F(h_i) > F(h_j)$) et la seconde une égalité stricte de la valeur du critère unique de synthèse ($F(h_i) = F(h_j)$). Ce résultat ne prend pas en compte certaines notions pourtant fréquemment rencontrées lorsque l'on cherche à comparer deux hypothèses concurrentes. En effet, les principales méthodes "classiques" d'agrégation s'appuient sur une exploitation directe des valeurs $F(h_i)$ et $F(h_j)$ pour comparer des hypothèses. La confiance "aveugle" apportée à ces valeurs ne tient aucunement compte de l'imprécision et de l'incertitude pouvant intervenir lors de l'évaluation d'hypothèses. De même, chercher à construire à tout prix un classement complet exige que toutes les hypothèses de l'ensemble H soient comparables entre elles. Or, dans de nombreux problèmes décisionnels, notre cas de contrôle y compris, il est souhaitable d'écarter des hypothèses jugées comme très peu pertinentes ou n'ayant pu être évaluées sur certains critères. De telles situations d'incomparabilité ne doivent pas être considérées comme des défauts de la procédure de décision mais bien comme des sources intéressantes de compréhension des hypothèses évaluées. Ainsi, statuer de l'incomparabilité d'une hypothèse parce qu'elle apparaît comme incohérente sur l'un des critères disponibles, permet d'identifier une interprétation qui n'aurait pas dû être générée et qui nécessiterait une révision des ressources linguistiques à l'origine de la génération de cette hypothèse.

Les approches par surclassement ont pour objectif de pallier la rigidité des méthodes "classiques" d'agrégation, notamment à travers la construction de relations floues de comparaison : les relations de surclassement. Vers la fin des années 1960, Bernard ROY[Roy, 1974] introduit le concept de surclassement pour comparer des hypothèses concurrentes. "une relation de surclassement est une relation binaire S définie dans H telle que $h_i S h_j$ si, étant donné ce que l'on sait des préférences du décideur et étant donné la qualité des évaluations des hypothèses et la nature du problème, il y a suffisamment d'arguments pour admettre que h_i est au moins aussi bonne que h_j, sans qu'il y ait de raison importante de refuser cette affirmation."

Comme nous allons le constater dans la section 3.1.3 consacrée à ces approches, les méthodes par surclassement permettent une exploitation de l'information apportée par chaque critère de comparaison disponible, mais également de prendre en compte l'imprécision, le compromis et l'incomparabilité pouvant intervenir lors d'un processus décisionnel basé sur de multiples critères hétérogènes.

Outre les perspectives d'obtenir à l'aide des approches par surclassement une méthodologie plus flexible, nous pouvons constater que les conditions d'application de ces méthodes correspondent avec beaucoup d'adéquation aux caractéristiques de nos tâches de contrôle du processus linguistique. En effet, Bernard Roy, en tant que fondateur de l'approche par surclassement, définit dans [Roy and Bouyssou, 1993] (chapitre V page 246) différentes conditions qui permettent de valider l'usage de ces méthodes. En insistant dans un premier temps sur l'intérêt que revêtent ces méthodes lorsque l'ensemble G des critères est au moins de cardinalité 3, quatre conditions sont dans un second temps exposées et il suffit que l'une d'elles soit remplie pour que l'usage d'une méthode d'AMCD par surclassement prenne tout son sens :

1. un des critères est défini sur une échelle qualifiée de "faussement quantitative". C'est-à-dire

que les écarts observés lors des comparaisons ne sont pas directement interprétables ;

2. les critères de l'ensemble G sont hétérogènes et donc pas forcément commensurables. C'est notre cas par exemple lorsque nous sommes en présence de critères numériques (statistiques) et binaires ;
3. le compromis recherché entre les critères, en d'autres termes la compensation d'une performance faible sur un critère par les autres critères utilisés, s'effectue selon des préférences établies par le décideur.
4. on souhaite intégrer des paramètres d'usage des critères, plus précisemment des seuils matérialisant l'imprécision des performances.

Les conditions d'application citées précédemment mettent bien en exergue la philosophie des approches par surclassement : proposer des outils méthodologiques applicables à des contextes décisionnels réels et non uniquement pour des problèmes rendus artificiels par des contraintes très fortes comme celle de la commensurabilité des critères.

3.2 Les approches par surclassement

Cette section décrit plus en détail les fondements et les concepts de base des approches par surclassement en AMCD. Parmi l'ensemble des méthodes proposées rentrant dans ce paradigme, nous verrons que notre choix s'est portée sur les méthodes apportant le plus de liberté au décideur dans la description du contexte décisionnel.

3.2.1 Modélisation des préférences et connaissances expertes

Nous avons énoncé précédemment que l'AMCD s'appuyait sur de nombreux domaines de recherche pour construire ses modèles décisionnels. Le domaine sous-jacent le plus présent, notamment dans les approches par surclassement, est celui de la modélisation des préférences.

Qu'il s'agisse d'une problématique de sélection, de classement ou de tri, la construction du résultat repose sur la comparaison des différentes hypothèses de l'ensemble H. Au delà de l'agrégation des performances atteintes sur les critères de l'ensemble G, la comparaison s'appuie sur la construction de relations binaires entre différentes paires d'hypothèses de l'ensemble H. En effet, lorsqu'un décideur se prononce sur un problème décisionnel, il s'appuie sur un système de préférences résultant d'un processus rationnel de comparaison. Ces préférences se matérialisent par des relations binaires établies (voir [Bouyssou and Vincke, 2006] pour une étude axiomatique approfondie de ces relations) entre deux hypothèses h_i et h_j de l'ensemble H et qui nous permettent de définir 4 relations entre les hypothèses :

- de préférence stricte $h_i P h_j$ (P : non-reflexive et asymétrique) ;
  Correspond à une situation où les performances obtenues par h_i sur les critères exploités justifient une préférence significative en faveur de h_i par rapport à h_j.
- d'indifférence $h_i I h_j$ (I : symétrique et réflexive) ;
  Désigne une situation où les différences de performances obtenues par h_i et h_j ne sont pas suffisament significatives pour établire une relation de préférence pour l'une ou l'autre des deux hypothèses comparées.
- de préférence faible $h_i Q h_j$ (Q : non-reflexive et asymétrique) ;
  Il s'agit du type de préférence le plus récemment introduit dans certaines méthodes par surclassement, afin de tenir compte d'une situation intermédiaire entre la préférence stricte et l'indifférence.

– d'incomparabilité $h_i R h_j$ (R : symétrique et irréflexible)).
Correspond à une situation où les performances conduisent à l'impossibilité de se prononcer sur la préférence ou l'indifférence entre deux hypothèses. Nous verrons précisément dans la section 3.3.1 dans quels cas cette incomparabilité apparaît et l'importance des connaissances qu'elle apporte sur notre problème.

La particularité des approches par surclassement réside principalement dans la prise en compte de l'ensemble de ces situations de comparaison, là où les approches plus "classiques" ne considèrent que les relations de préférence et d'indifférence. De plus, les approches par surclassement considèrent que les différentes situations de préférence décrites précédemment ne sont pas forcément transitives. Cette propriété, bien qu'à l'origine d'une complexité opératoire importante lors de la comparaison des hypothèses, permet d'obtenir des résultats moins "artificiels" et plus conforme à la réalité du contexte décisionnel.

L'ensemble des relations binaires établies entre les hypothèses H forme une structure de préférences (toute paire d'hypothèses est reliée par une seule de ces relations). L'automatisation partielle du processus de décision que nous cherchons à conduire vise à construire de telles structures de préférences. Parmi les différentes situations de préférence pouvant être établies entre des hypothèses concurrentes, une comparaison effectuée uniquement sur la base des vecteurs de performances n'offre qu'une information relativement pauvre. En effet, seules des relations de dominance Δ peuvent être établies entre les éléments de l'ensemble H si les vecteurs de performances sont les seules connaissances exploitées. Une relation de dominance est établie entre deux hypothèses si tous les critères de comparaison qu'elles ont en commun s'accordent pour désigner une hypothèse comme meilleure :

$$h_i \Delta h_j \Leftrightarrow \begin{cases} \forall g_k \in G, g_k(h_i) \geq g_k(h_j) \\ \exists g_k \in G, g_k(h_i) > g_k(h_j) \end{cases}$$

Ainsi, pour enrichir cette exploitation des vecteurs de performances en vue de construire des relations binaires entre les hypothèses, les méthodes d'AMCD par surclassement ont recours à l'intégration d'un autre type de préférences. Contrairement aux préférences décrites précédemment qui caractérisent des relations entre les hypothèses, ces préférences, également appelées informations préférentielles, constituent des paramètres décisionnels décrivant la façon dont les vecteurs de performances doivent être agrégés. En tant que paramètres définis *a priori*, ces informations représentent des connaissances qu'un expert du domaine (le décideur) peut introduire dans le processus décisionnel. De même que les probabilités, les fonctions d'appartenance ou de croyance, ces préférences permettent d'introduire dans la procédure d'agrégation une mesure de compensation, d'imprécision, d'incertitude et d'incomparabilité lié à l'évaluation des hypothèses. Dans une approche d'AMCD, ces différentes mesures se matérialisent par un ensemble de paramètres associés aux critère $G : \{g_1, g_2, ..., g_m\}$, tels que :

– un ensemble de poids d'importance $W : \{w_1, w_2, ..., w_m\}$;
détermine la contribution d'un critère dans l'évaluation des hypothèses et détermine également les mesures de compensation inter-critères.
– un ensemble de seuils de préférence $P : \{p_1, p_2, ..., p_m\}$;
p_k représente la plus petite différence $g_k(h_i) - g_k(h_j)$ traduisant une situation de préférence en faveur de h_i par rapport à h_j sur le critère g_k .
– un ensemble de seuils d'indifférence $Q : \{q_1, q_2, ..., q_m\}$;
q_k représente la plus grande différence $g_k(h_i) - g_k(h_j)$ préservant une situation d'indifférence entre h_i et h_j sur le critère g_k .

– un ensemble de seuils veto $V : \{v_1, v_2, ..., v_m\}$.
où v_k représente la plus petite différence $g_k(h_i) - g_k(h_j)$ tolérée pour que h_i et h_j restent comparables. Un tel seuil permet par exemple de considérer comme incomparable une hypothèse ayant une performance trop faible sur un critère pour qu'elle puisse être compensée par les autres critères.

Pour compléter l'information de performance portée par les critères, le décideur a la possibilité de leur associer ces différents paramètres techniques, passant ainsi du statut de *vrai-critère* (sans seuil), à celui de *pseudo-critère* (avec deux seuils, de préférence et d'indifférence), en passant par les *quasi-critères* (avec un seuil). Ainsi, en fonction des informations préférentielles associées aux critères, on sera en mesure d'établir différents types de structures de préférences :

– un préordre total avec des *vrai-critères* (classement sans ex-aequos) ;
– un quasi-ordre avec des *quasi-critères* (classement avec ex-aequos) ;
– un pseudo-ordre avec des *pseudo-critères* (classement avec ex-aequos et incomparabilités).

La construction d'une structure de préférence repose alors sur l'évaluation des vecteurs de performances tout en respectant les conditions de comparaison stipulées par les paramètres décisionnels. Nous allons voir dans la fin de cette section comment les approches par surclassement, dans une démarche prescriptive, construisent des relations binaires de comparaison entre les hypothèses en exploitant l'ensemble des informations disponibles (performances et paramètres décisionnels).

3.2.2 Le surclassement comme relation générique de comparaison

La comparaison des différents éléments de l'ensemble H des hypothèses repose sur la construction d'un seul type de relation, les relations de surclassement, qui permettent cependant d'exprimer les différentes situations présentées précédemment, à savoir la préférence, l'indifférence ou l'incomparabilité.

– $h_i P h_j$ si $h_i S h_j$ et non $h_j S h_i$
– $h_i I h_j$ si $h_i S h_j$ et $h_j S h_i$
– $h_i R h_j$ si non $h_i S h_j$ et non $h_j S h_i$

La construction d'une relation de surclassement $h_i S h_j$ repose sur deux notions : la concordance et la discordance. La `concordance` représente et quantifie la propension des critères à valider l'assertion de surclassement. Il est relativement intuitif de transposer cette notion au paradigme du vote et de la théorie du choix social, où chaque critère correspond à un votant ou un groupe de votants (dont la dimension dépend du poids associé au critère), et où la concordance quantifie la proportion des votants acceptant la supériorité de h_i sur h_j.

La `discordance` représente et quantifie la propension des critères à refuser l'assertion de surclassement de h_i sur h_j. Toujours en se rapportant au paradigme du vote, elle correspond à l'opposition. Cette prise en compte de la discordance est primordiale, car elle correspond à une caractéristique forte des approches par surclassement, celle de considérer lors de la comparaison tous les critères disponibles. De nombreuses méthodes dites multicritères telles que les ordres lexicographiques n'exploitent au final qu'un sous-ensemble, voir un seul critère dit dictateur. Il suffit alors que ce critère dictateur soit en faveur d'une hypothèse pour qu'une relation de préférence soit établie. Comme nous allons le voir à travers une description détaillée des aspects méthodologiques d'une approche par surclassement (section 3.2.3), la discordance exploite les seuils veto définis dans les paramètres décisionnels pour soit conduire à une situation d'incomparabilité, soit réduire l'importance de la concordance lors de la comparaison de deux hypothèses.

Ainsi, pour résumer ces différentes notions, une relation de surclassement entre deux hypothèses est établie si les deux conditions suivantes sont remplies :

1. si une majorité suffisante de critères valides s'accorde pour établir le surclassement (concordance) ;
2. et si l'opposition à cette relation n'est pas trop forte (non discordance).

Nous allons également constater dans la section 3.2.3 que la construction de relations de surclassement dépend principalement des vecteurs de performances associés aux hypothèses comparées, mais également des différents paramètres décisionnels définissant le contexte de la prise de décision.

Qu'il s'agisse de relation de préférence stricte/faible, d'indifférence, d'incomparabilité ou de surclassement, on constate aisément qu'une structure de préférence, en tant que regroupement des relations binaires établies sur l'ensemble H, ne constitue pas une réponse directement exploitable pour une problématique de classement, sélection ou de tri. Une méthode d'AMCD par surclassement est donc définie par deux étapes successives :

1. la construction des relations de surclassement formant la structure de préférences ;
2. l'exploitation de la structure de préférences en vue de construire un classement des hypothèses, d'extraire un sous-ensemble de meilleures hypothèses ou de les trier.

De nombreuses méthodes s'appuyant sur les relations de surclassement ont été développées lors des cinq dernières décennies [Guitouni and Martel, 1998]. Ces méthodes se différencient à la fois par la façon dont les relations de surclassement sont construites et par la manière dont elles sont interprétées pour répondre à une problématique visée.

Parmi ces méthodes, les plus connues et utilisées sont les méthodes `ELECTRE` [Bouyssou, 2001] développées par Bernard ROY, mais la méthode `PROMETHEE` [Brans and Vincke, 1985], de par la liberté qu'elle accorde au décideur pour la modélisation de ses préférences, est également très répandue. Nous pouvons également citer la méthode `MELCHIOR` [Leclercq, 1984] qui a l'originalité de prendre en compte des préférences non quantitatives lors de la construction des relations entre hypothèses.

Les méthodes `MELCHIOR` et `PROMETHEE` sont cependant dédiées à une unique problématique, celle du rangement, alors que notre objectif est de disposer d'une base méthodologique commune pour les trois problématiques envisagées dans le cadre du contrôle d'un processus de TALN : le rangement, la sélection ou le tri. Comme nous allons le constater par la suite, les méthodes `ELECTRE` (dans leurs versions `I-III` et `TRI`) nous permettent d'envisager cette généralisation autour d'une unique procédure de construction des relations de surclassement. De plus, ces méthodes apparaissent comme des références dans le domaine de l'AMCD et sont donc à l'origine de nombreux travaux attestant de leur efficacité à proposer des recommandations pour des problèmes décisionnels concrets.

Nous nous sommes donc plus particulièrement intéressé aux différentes méthodes `ELECTRE` : `ELECTRE I-II-III-IV`, et `ELECTRE TRI`. Les quatres versions d'`ELECTRE` se différencient par les problématiques qu'elles cherchent à résoudre (le choix pour les versions I et le classement pour les versions II, III et IV et le tri pour `ELECTRE TRI`), mais également par le type des critères pris en compte (vrai, quasi ou pseudo critères) et la façon dont les relations de surclassement sont construites. En effet, les progrès en terme de modélisation des préférences ont fait évoluer les

méthodes ELECTRE, notamment afin qu'elles puissent prendre en compte des critères avec seuils de préférence et d'indifférence.

Initialement conçue pour une problématique de rangement, la méthode ELECTRE III nous est apparue comme la plus prometteuse pour construire notre méthodologie générique de contrôle. En effet, les différentes informations préférentielles sur l'utilisation des critères sont optionnelles. La méthode ELECTRE III peut donc apparaître sous sa forme la plus complète en agrégant des pseudo-critères, mais peut également se rapprocher de méthodes plus simples comme ELECTRE I lorsqu'il s'agit de vrais-critères.

3.2.3 ELECTRE III et ELECTRE TRI

Construction des relations de surclassement

Outre la particularité de prendre en compte des pseudo-critères, c'est-à-dire des informations préférentielles détaillées, les relations de surclassement construites par la méthode ELECTRE III sont associées à un degré de surclassement, qui permet de quantifier la crédibilité à accorder à la relation construite. Ce degré de surclassement est compris entre 0 et 1, il vaut 1 si tous les critères sont concordants et 0 si la discordance est maximale, c'est-à-dire qu'au moins un des critères a posé son veto envers une assertion de surclassement.

La méthode ELECTRE III construit entre chaque paire d'hypothèses $h_i, h_j \in H$ des relations de surclassement $h_i S h_j$ qualifiées par un degré de crédibilité de surclassement $S(h_i, h_j) \in [0, 1]$. Cet indice est calculé à partir de deux mesures représentant la concordance $C(h_i, h_j)$ et la discordance $D(h_i, h_j)$.

$$C(h_i, h_j) = \frac{1}{P}.\sum_{k=1}^{m} w_k.c_k(h_i, h_j) \text{ où } P = \sum_{k=1}^{m} w_k$$

Le calcul de cet indice de concordance global s'appuie sur des indices de concordance partiels $c_k(h_i, h_j)$ établis pour chaque critère :

$$c_k(h_i, h_j) = \begin{cases} 0, & \text{si } g_k(h_j) - g_k(h_i) \geq p_k \\ \frac{p_k - (g_k(h_i) - g_k(h_j))}{p_k - q_k}, & \text{si } q_k \leq g_k(h_j) - g_k(h_i) \leq p_k \\ 1, & \text{si } g_k(h_j) - g_k(h_i) \leq q_k \end{cases}$$

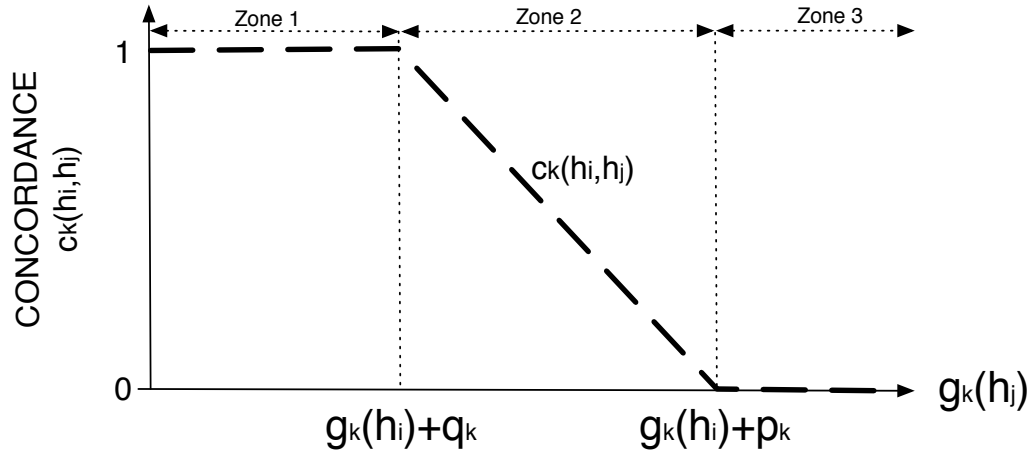

FIG. 3.1 – Illustration graphique de la concordance sur le critère k avec l'assertion $h_i S h_j$

La figure 3.1 illustre graphiquement l'évaluation de la concordance entre deux hypothèses h_i et h_j sur un critère g_k, où la performance de l'hypothèse h_j sur le critère g_k est matérialisée par l'axe des abscisses. La zone 1 correspond à une situation de concordance maximale $c_k(h_i, h_j) = 1$ car $g_k(h_j) < g_k(h_i) + q_k$, c'est-à-dire que l'hypothèse h_i est considérée comme au moins aussi pertinente que h_j sur le critère k, en tenant compte de l'imprécision introduite par le seuil d'indifférence q_k. La zone 3 correspond à un cas de concordance nul $c_k(h_i, h_j) = 0$, car la performance obtenue par h_j est largement supérieure ($g_k(h_j) > g_k(h_i) + p_k$) à celle obtenue par h_i sur le critère k.

La zone 2 correspond à une situation de concordance intermédiaire $c_k(h_i, h_j) \in]0, 1[$, où la différence $q_k < g_k(h_j) - g_k(h_i) < p_k$.

La discordance est quantifiée à partir d'indices de discordance partiels $d_k(h_i, h_j)$ tels que :

$$d_k(h_i, h_j) = \begin{cases} 1, & \text{si } g_k(h_j) - g_k(h_i) \geq v_k \\ \frac{g_k(h_j) - g_k(h_i) - p_k}{v_k - p_k} & \text{si } p_k < g_k(h_j) - g_k(h_i) < v_k \\ 0, & \text{si } g_k(h_j) - g_k(h_i) \leq p_k \end{cases}$$

L'indice de discordance partiel sur un critère k entre h_i et h_j ($d_k(h_i, h_j)$) vaut 1 (zone 3 Fig. 3.2) lorsque la différence de performance $g_k(h_j) - g_k(h_i)$ est supérieure au seuil veto associé au critère k.

L'indice de concordance $C(h_i, h_j)$ et les indices de discordance partiels $d_k(h_i, h_j)$ sont ensuite utilisés pour calculer le degré de crédibilité $\sigma(h_i, h_j)$ final accordé à la relation de surclassement $h_i S h_j$. Soit F les indices des critères $G : \{g_1, g_2, ..., g_m\}$ ($F : \{1, 2, ..., m\}$) :

$$\sigma(h_i, h_j) = C(h_i, h_j) \prod_{k \in \overline{F}} \frac{1 - d_k(h_i, h_j)}{1 - C(h_i, h_j)} \text{où } \overline{F} = \{k \in F / d_k(h_i, h_j) > C(h_i, h_j)\}$$

À partir de la formule précédente, on remarque aisément que l'indice de crédibilité $\sigma(h_i, h_j)$ est égal à l'indice de concordance $C(h_i, h_j)$ si aucun des critères utilisés n'est discordant avec l'assertion de surclassement. Au contraire, cet indice est nul si un des critères est complètement

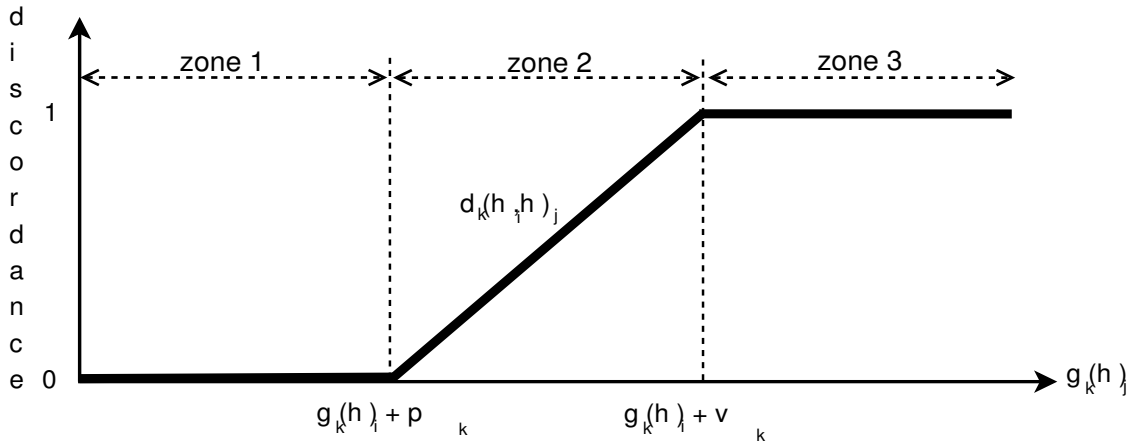

FIG. 3.2 – Illustration graphique de la discordance sur le critère k avec l'assertion $h_i S h_j$

discordant. L'indice de crédibilité $\sigma(h_i, h_j)$ est ensuite comparé à un seuil de coupe $\alpha \in [0, 1]$ qui détermine la limite d'acceptabilité d'une relation de surclassement :

$$h_i S h_j \text{ si } \sigma(h_i, h_j) \geq \alpha$$

Interprétation des relations de surclassement pour une problématique de classement

La méthode `ELECTRE III` (annexe A.4) a été conçue initialement pour répondre à une problématique de classement. La deuxième étape de cette méthode, dite d'interprétation, vise à construire un préordre partiel (classement avec *ex-aequo* et prise en compte de l'incomparabilité) à partir de la structure de préférences construite lors de la première étape.

L'algorithme de construction de ce préordre partiel, que nous présentons ci-dessous, peut apparaître comme complexe dans la mesure où ce préordre est établi à partir de deux préordres totaux (classement avec *ex-aequo* sans incomparabilité), l'un ascendant et l'autre descendant. Cette construction intermédiaire a pour objectif principal d'identifier les hypothèses incomparables dont l'intégration entre deux autres hypothèses semble délicat. La construction du préordre partiel médian, en tant qu'intersection des relations de préférence et d'indifférence des deux préordres totaux, permet d'identifier ces cas particuliers où une attention particulière devrait être apportée sur le calcul de ses performances.

La construction des préordres totaux repose sur un paramètre technique supplémentaire appelé seuil de discrimination des indices de crédibilité $s(\lambda)$, où λ correspond à l'indice de crédibilité maximal des relations de surclassement : $\lambda = max_{h_i, h_j \in H} S(h_i, h_j)$. Le seuil de discrimination $s(\lambda)$ qui s'exprime fréquemment sous la forme $s(\lambda) = \alpha + \beta.\lambda$ permet d'effectuer une sélection des arcs (i.e. des relations de surclassement) dont l'indice de crédibilité est proche de l'indice maximum λ. Les valeurs des variables α et β peuvent varier afin d'obtenir plus de tolérance lors de la comparaison des indices de crédibilité. Nous les avons cependant fixées à $\alpha = 0,30$ et $\beta = -0,15$, ce qui correspond à des valeurs fréquemment utilisées ([Martin and Legret, 2005] et [Roy and Bouyssou, 1993] page ...). Le seuil de discrimination $s(\lambda)$ permet d'identifier les relations de surclassement les plus significatives, c'est-à-dire celles répondant aux spécifications suivantes :

$$S(h_i, h_j) \geq \lambda - s(\lambda) \text{ et } |S(h_i, h_j) - S(h_j, h_i)| \leq s(\lambda)$$

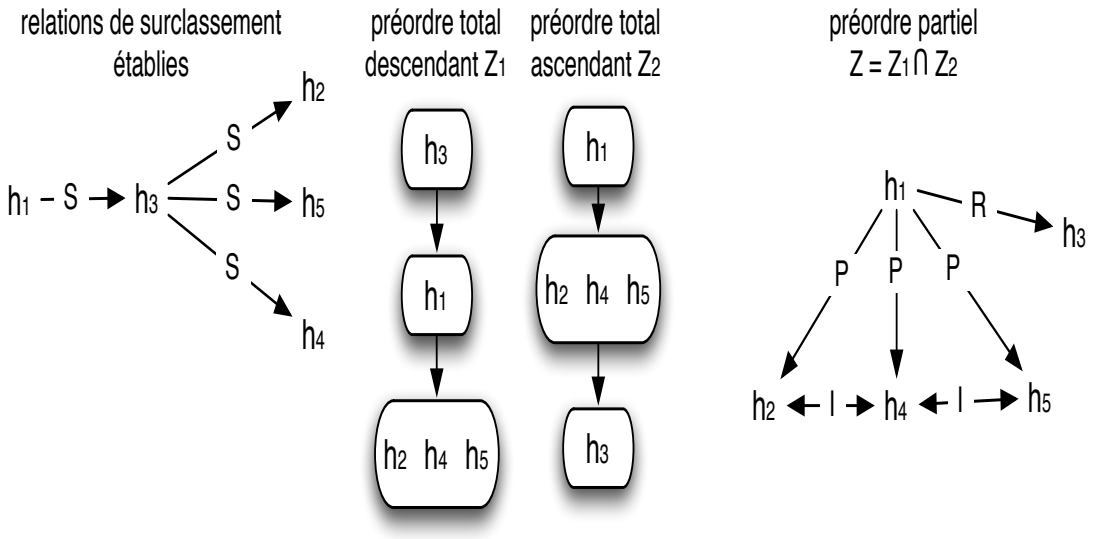

FIG. 3.3 – Interprétation des relations de surclassement établies par `ELECTRE III` en vue de répondre à une problématique de classement, illustration de la prise en compte de l'incomparabilité entre h_1 et h_3.

À partir de ce sous-ensemble de relations qui ne sont désormais plus valuées, on calcule pour chacune des hypothèses h_i sa qualification $Q(h_i)$ qui correspond à la différence entre le nombre d'hypothèses surclassées par h_i et le nombre d'hypothèses qui la surclassent. On établit ensuite un premier distillat D_1 correspondant aux hypothèses ayant une qualification maximale $D_1 : h_k$ tel que $Q(h_k) = max_{h_i} Q(h_i)$. Cet ensemble D_1 consitue le premier élément du préordre total descendant et la procédure se poursuit dans H/D_1.

Ce même algorithme est ensuite reconduit sur l'ensemble H, mais ne retenant lors de la construction de D_1 que les hypothèses ayant la plus faible qualification. Ce deuxième préordre total est qualifié d'ascendant, partant des hypothèses les moins pertinentes pour aller vers les hypothèses les plus pertinentes.

À partir de ces deux préordres totaux, un préordre partiel médian est établi. Il regroupe l'ensemble des relations d'indifférence (hypothèses appartenant à un même distillat), de préférence (hypothèses appartenant à des distillats de rang différents) et d'incomparabilité lorsque les deux préordres divergent sur le classement d'hypothèses.

`ELECTRE TRI` : une variante d'`ELECTRE III` pour les problématiques de tri

`ELECTRE TRI` (voir annexe A.5) est une méthode s'inspirant de `ELECTRE III` visant à répondre à une problématique de tri (affectation des hypothèses dans des classes définies *a priori*). Elle reprend en effet les méthodes de calcul des indices de concordance, de discordance et de crédibilité pour affecter les différentes hypothèses de l'ensemble H dans des classes prédéfinies $C : c_1, c_2, ..., c_t$. Ainsi, au lieu d'obtenir un classement complet ou partiel des hypothèses, la méthode `ELECTRE TRI` procède à des regroupements d'hypothèses de pertinence comparable dans les

différentes classes définies *a priori*. Une classe est alors définie par un ensemble de performances minimales, qui permettent de déterminer les seuils d'acception des hypothèses dans la classe. On compare donc les vecteurs de performances des hypothèses avec les limites d'acceptabilité des différentes classes.

Afin d'éviter une confusion relativement fréquente et pour mieux comprendre cette méthode, il apparaît important de rappeler la distinction entre les notions de tri et de classification. Bien que ces deux problématiques visent à obtenir un regroupement d'hypothèses partageant des similarités (de pertinence dans notre cas), le tri impose que les différentes classes suivent un ordre d'importance. Ainsi, les hypothèses regroupées au sein de la classe c_{n-1} seront moins pertinentes que les hypothèses de la classe c_n.

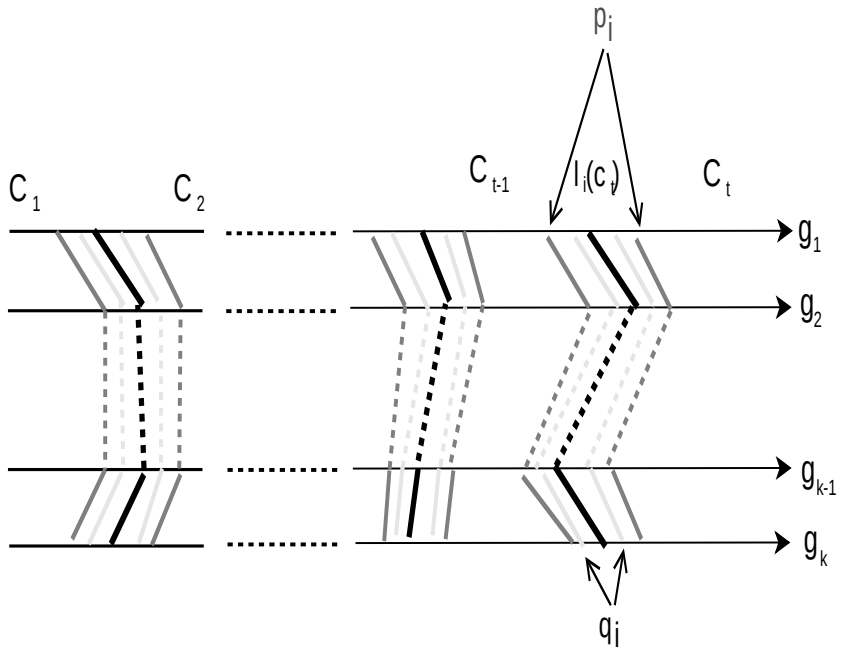

FIG. 3.4 – Affectation des hypothèses aux classes définies *a priori* en utilisant la méthode `ELECTRE TRI`

`ELECTRE TRI` se différentie donc de `ELECTRE III` uniquement par le fait que les relations de surclassement ne sont pas établies entre les hypothèses de l'ensemble H mais entre les hypothèses et les classes C. La décision d'affecter une hypothèse dans une des classes peut s'effectuer selon deux approches, l'une optimiste, l'autre pessimiste. La méthode pessimiste suit l'algorithme

suivant :

Précondition :
$C : \{c_t, c_{t-1}, ..., c_2, c_1\}$: classes définies *a priori*;
$H : \{h_1, h_2, ..., h_n\}$: hypothèses à trier;
Postcondition :
$\forall h_i \in H, \exists c_j \in C$ tel que $h_i \in c_j$;
Boucle :

```
for h_i ∈ H do
  for c_j ∈ {c_t, c_{t-1}, ..., c_2, c_1} do
    if h_i S c_j > α then
      h_i ∈ c_j
    end if
  end for
end for
```

Algorithme 1 : Algorithme de tri réalisé par `ELECTRE TRI` : version pessimiste

alors que la méthode optimiste évalue le surclassement des hypothèses par rapport aux classes dans le sens inverse :

Précondition :
$C : \{c_1, c_2, ..., c_{t-1}, c_t\}$: classes définies *a priori*;
$H : \{h_1, h_2, ..., h_n\}$: hypothèses à trier;
Postcondition :
$\forall h_i \in H, \exists c_j \in C$ tel que $h_i \in c_j$;
Boucle :

```
for h_i ∈ H do
  for c_j ∈ {c_t, c_{t-1}, ..., c_2, c_1} do
    if h_i S c_j > α then
      h_i ∈ c_j
    end if
  end for
end for
```

Algorithme 2 : Algorithme de tri réalisé par `ELECTRE TRI` : version optimiste

L'affectation des hypothèses dans les différentes classes nécessite de disposer d'éléments de comparaison pour chacun des critères qualifiant la pertinence des hypothèses. Cette comparaison s'appuie sur de nouveaux paramètres pouvant également s'apparenter à des informations préférentielles, que l'on désigne sous le terme de vecteur d'acceptabilité d'une classe (figure 3.4). Ainsi chaque classe c_j est associée à un vecteur $L : \{l_1, l_2, ...l_m\}$, où $l_k(c_j)$ correspond à la limite d'acceptabilité de la classe c_j pour le critère d'indice k. Ainsi, tout en conservant les diverses informations préférentielles (seuils de préférence, indifférence ou veto), les relations de surclassement sont établies à l'aide des formules présentées précédemment en comparant le vecteur de performances de chaque hypothèse et le vecteur des limites d'acceptabilité associées aux classes.

Contrairement à la méthode `ELECTRE III`, les relations de surclassement établies par la méthode `ELECTRE TRI` n'ont pas besoin d'être interprétées pour répondre à la problématique visée. En effet, le résultat du tri est directement obtenu lors de la construction des relations de surclassement. On constate donc que la version d'`ELECTRE` dédiée aux problématiques de tri est moins complexe (linéaire en fonction du nombre de classes, d'hypothèses et de critères de comparaison)

car elle ne repose pas sur une comparaison deux à deux des hypothèses de l'ensemble H et car elle ne nécessite pas de méthode supplémentaire d'interprétation des relations de surclassement construites.

En appliquant l'un des deux algorithmes présentés, version pessimiste 4 ou optimiste 5, on obtient directement une affectation de chaque hypothèse dans les classes définies *a priori*. Dans un même esprit que la construction d'un classement à partir de deux pré-ordres complets antagonistes effectuée par la méthode `ELECTRE III`, les deux algorithmes de tri peuvent être appliqués et leurs résultats comparés pour identifier les situations d'incomparabilité entre des hypothèses et les vecteurs d'acceptabilité des classes.

Interprétation des relations de surclassement pour une problématique de sélection

Nous venons de constater que les méthodes `ELECTRE III` et `ELECTRE TRI`, dédiées respectivement aux problématiques de classement et de tri, disposaient d'une base méthodologique commune de comparaison des hypothèses soit entre elles soit vis-à-vis de profils de classes. Dans la section 2.1.1 consacrée à la définition de la notion de contrôle d'un processus de TALN, nous avons vu que les problématiques visées concernaient le classement, le tri mais également la sélection. Nous rappelons que l'objectif d'un contrôle par sélection est d'identifier un sous-ensemble d'hypothèses le plus restreint possible, que l'on peut qualifier de meilleures, c'est-à-dire plus pertinentes que les autres.

La première version des méthodes `ELECTRE` est dédiée à cette problématique, mais contrairement aux deux méthodes présentées précédemment, elle ne repose pas sur la même procédure de construction des relations de surclassement. `ELECTRE I` est une méthode plus simple et donc plus répandue. La simplicité de cette méthode provient du fait qu'elle exploite des vrais-critères pour construire des relations de surclassement entre les différentes hypothèses concurrentes. Il n'est donc pas possible de faire intervenir des marges d'imprécision et d'incertitude lors de la comparaison des hypothèses. De manière analogue à la méthode `ELECTRE III`, `ELECTRE I` construit dans un premier temps une structure de préférences (figure 3.5) regroupant l'ensemble des relations de surclassement établies entre les hypothèses concurrentes.

Cette structure de préférences est ensuite interprétée pour identifier le sous-ensemble d'hypothèses jugées comme les plus pertinentes. Une structure de préférences constitue un graphe, dont les sommets correspondent aux hypothèses comparées et les arêtes reprennent les relations de surclassement établies entre les hypothèses.

La procédure d'identification du sous-ensemble d'hypothèses privilégiées s'appuie sur des méthodes issues de la théorie des graphes. En effet, ce sous-ensemble correspond au noyau du graphe. Nous rappelons qu'un noyau d'un graphe correspond à un ensemble de sommets stable et dominant, c'est-à-dire dont les sommets ne sont pas joints deux à deux et où chaque sommet hors du noyau est relié à un élément du noyau.

La structure de préférences résultant de la comparaison des hypothèses correspond à un graphe possédant éventuellement des circuits (succession de sommets dont le départ est équivalent à l'arrivée), et l'identification d'un noyau dans ce genre de graphe constitue un problème `NP-complet` [Chvatal, 1973].

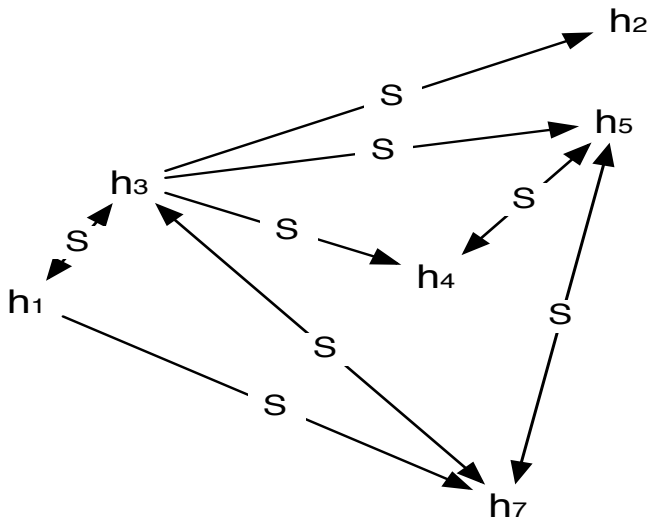

FIG. 3.5 – Structure de préférences regroupant les relations de surclassement établies entre les hypothèses concurrentes

C'est pourquoi la procédure de sélection procède tout d'abord à une réduction des circuits maximaux de la structure de préférences. L'ensemble des circuits maximaux disjoints est noté $H_\sigma : \{\sigma_1, \sigma_2, ..., \sigma_p\}$. Afin de former un graphe orienté sans circuit, une relation notée $\succ_\sigma$ est construite entre les différents circuits σ_i en appliquant la règle suivante :

$$\forall \sigma_i, \sigma_j \in H_\sigma : \sigma_i \succ_\sigma \sigma_j \Leftrightarrow \sigma_i \neq \sigma_j \text{ et } \exists h_i \in \sigma_i \text{ et } \exists h_j \in \sigma_j \text{ tels que } h_i S h_j$$

Le graphe construit autour de cette relation asymétrique et sans circuit admet un unique noyau $N \in H_\sigma$ beaucoup plus simple à identifier. Ce noyau N est construit itérativement de la façon suivante. On identifie tout d'abord un ensemble N_1 d'éléments de H_σ qui ne sont pas préférés par un autre élément de H_σ selon $\succ_\sigma$. On supprime de H_σ les éléments de N_1 et les éléments de H_σ / N_1 tels qu'il existe un élément de N_1 qui lui soit préféré. Si le sous-ensemble restant H_σ^1 n'est pas vide, on lui applique les étapes précédentes. Lorsque H_σ^k est vide, le noyau correspond alors à :

$$N = N_1 \bigcup N_2 \bigcup ... \bigcup N_k$$

Si N est composé de singletons, on peut poser que $H\prime = N$. Si au contraire N contient un élément σ qui est un circuit, alors la sélection d'une hypothèse nécessite une étude plus approfondie de ce circuit afin d'identifier l'élément qui semblerait prévaloir. Nous verrons dans la section 3.3 que les particularités de notre contexte nous permettent de nous affranchir de cette étude.

Pour une description plus détaillée de cet algorithme d'interprétation d'une structure de préférences en vue de répondre à une problématique de sélection, le lecteur intéressé pourra consulter les pages 366-368 de [Roy and Bouyssou, 1993].

Exemple

La figure 3.6 illustre graphiquement cette procédure de sélection sur un exemple extrait de [Roy and Bouyssou, 1993] pages 366-367.

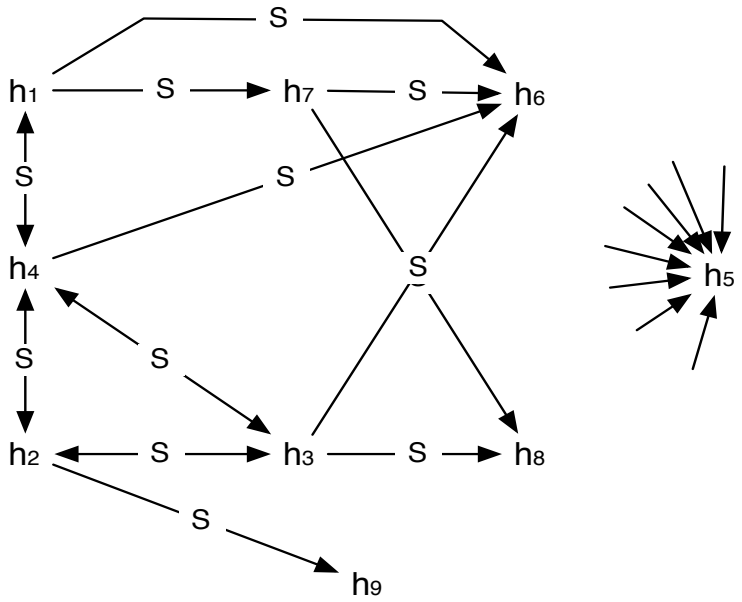

FIG. 3.6 – Représentation sous forme de graphe d'une structure de préférences, où un arc orienté de h_i vers h_j signifie $h_i S h_j$. L'absence d'arc dénote une situation d'incomparabilité. h_5 est surclassée par toutes les autres hypothèses.

La première étape de la procédure de sélection repose sur l'identification de tous les circuits maximaux : $\Sigma : \{\sigma_1, \sigma_5, \sigma_6, \sigma_7, \sigma_8, \sigma_9\}$, où $\sigma_1 = \{ h_1, h_2, h_3, h_4\}$ et $\sigma_i = h_i$ pour $i \in \{5, 6, 7, 8, 9\}$. La construction des $\succ_\sigma$ relations vérifie que $\sigma_1 \succ_\sigma \sigma_i \forall i \neq 1$. Le noyau identifié est alors $N = \{\sigma_1\}$.

La procédure d'interprétation que nous venons de décrire et illustrer sur un exemple exploite la propriété de graphe orienté des structures de préférences pour répondre à cette problématique de sélection. Les structures de préférences construites par la méthodes `ELECTRE III` correspondent également à des graphes orientés qui ont la propriété supplémentaire d'être valués par les indices de crédibilité. Afin de disposer d'une procédure générique de comparaison pour les trois problématiques envisagées dans notre approche de contrôle, nous proposons d'appliquer l'algorithme de sélection sur les structures de préférences générées par la méthode `ELECTRE III` (section 3.2.3).

Ceci nous permet donc de disposer d'une méthode commune de comparaison pour répondre aux trois problématiques de tri, de classement et de sélection.

Il est évident que nous aurions pu également exploiter le classement établi par la procédure dédiée à cette problématique (section 3.2.3) pour identifier un sous-ensemble d'hypothèses à privilégier. L'ensemble des hypothèses établi en tête de classement peut en effet constituer une réponse à cette problématique de sélection. Cependant, lorsque l'ensemble initial d'hypothèses concurrentes est très volumineux, il n'est pas pertinent en terme de complexité de procéder à un classement complet pour identifier un sous-ensembles d'hypothèses privilégiées, dans ce cas, il est préférable d'exploiter la méthode d'interprétation issue d'`ELECTRE I`.

3.3 Une approche par surclassement pour le contrôle des indéterminations

Dans le cadre de la mise en place d'une méthodologie générique de contrôle des cas d'indétermination apparaissant lors du processus d'analyse linguistique, nous nous sommes donc rapproché de l'AMCD et plus précisément des approches par surclassement. Lors de la présentation de ces approches (section 3.1.3), nous avons constaté que notre contexte d'application remplissait les conditions d'usage de ces méthodes. Cette section est focalisée sur la transposition de ces méthodes vers notre tâche de contrôle.

3.3.1 Particularités du contexte décisionnel

Les méthodes d'AMCD ont principalement été définies pour constituer un ensemble d'outils méthodologiques guidant un (ou des) décideur(s) et un homme d'étude vers l'élaboration d'une stratégie décisionnelle. Comme nous allons le constater plus clairement dans le chapitre suivant, nos travaux visent à produire des outils informatiques directement exploitables par un décideur (linguiste ou informaticien en charge du paramétrage de `TiLT`) en vue du contrôle d'un cas d'indétermination. Cette informatisation des méthodes d'AMCD n'est pas une originalité de nos travaux, dans la mesure où des logiciels tels que `ELECTRE TRI 2.0a` [Mousseau *et al.*, 1999], `ELECTRE IS`, `ELECTRE III-IV, version 3.x` [Skalka *et al.*, 1994], etc, ont déjà été proposés pour automatiser une prise de décision. Ces logiciels ne visent pas à se substituer à un homme d'étude, mais à fournir une interface d'aide à la mise en place d'une stratégie décisionnelle.

Les outils informatiques que nous avons ainsi développés (section 4) constituent alors des interfaces dédiées à la mise en place de profils d'analyse de `TiLT` et plus particulièrement aux sections spécialisées sur les étapes de contrôle.

De plus, la plupart des travaux scientifiques impliqués dans une démarche d'AMCD présente l'utilisation d'une méthode pour une problématique particulière de décision : l'implantation d'une centrale, d'un autoroute, le classement de projets, etc. Nos objectifs ne s'insèrent pas dans un tel cadre décisionnel complètement défini et figé. Nous cherchons en effet à proposer une base méthodologique et des outils pouvant répondre aux différents cas d'indétermination qui apparaissent lors d'un processus d'analyse linguistique. La méthode et les outils associés (interface de configuration) doivent donc permettre à un décideur d'introduire ses connaissances et intuitions sur le contexte décisionnel, mais également lui permettre de l'adapter à des particularités du contexte d'application et d'évaluer ces connaissances empiriques.

Cette dernière considération est d'ailleurs à l'origine de notre démarche consistant à relier l'AMCD au TALN. En effet, l'utilisation d'un même processus d'analyse linguistique sur différents types de corpus nécessite une adaptation des ressources linguistiques utilisées ainsi que des stratégies d'analyses et donc potentiellement des stratégies de contrôle. Cette adaptation peut dans de nombreux cas reposer uniquement sur les connaissances et intuitions d'un expert (décideur). Contrairement aux approches entièrement basées sur l'usage de corpus d'apprentissage, les méthodes par surclassement permettent à travers la construction d'une famille de critères et des informations préférentielles associées d'intégrer cette connaissance. Nous verrons dans la section suivante que cette propriété a énormément influencé nos choix en termes de modélisation et d'implémentation.

3.3.2 Une méthodologique axée sur la généricité

Nous avons vu précédemment qu'une approche par surclassement pouvait répondre aux problématiques de classement (section 3.2.3), sélection (section 3.2.3) et de tri (section 3.2.3). Bien que les développements méthodologiques en AMCD se soient orientés vers la mise en place de méthodes dédiées à une problématique et à un contexte décisionnel donnés[8], nous proposons d'exploiter leur base commune, c'est-à-dire la construction de relations de surclassement, pour répondre aux différentes problématiques. Nous avons en effet vu que la méthode `ELECTRE TRI` se différenciait d'`ELECTRE III` par le fait que les hypothèses ne se comparaient pas entre elles mais vis-à-vis de limites d'acceptabilité, tout en conservant cependant une méthode commune de construction des relations de surclassement.

Nous avons également constaté que la méthode `ELECTRE I`, méthode dédiée à la problématique de sélection, s'appuyait sur une structure de préférences similaire à celle construite par `ELECTRE III` pour identifier l'ensemble des meilleures hypothèses. En effet, la structure de préférences construite par `ELECTRE III` se différencie de celle construite par `ELECTRE I` par la quantification des relations de surclassement à l'aide d'indices de crédibilité. Ainsi, la phase d'interprétation visant à extraire un sous-ensemble de meilleures hypothèses à partir de la structure de préférences construite par `ELECTRE I` semble tout à fait applicable aux structures de préférences construites par `ELECTRE III`.

Nous avons vu dans la section 3.2.2 que la notion de surclassement apparaissait comme une relation générique de comparaison d'hypothèses concurrentes. Nous poursuivons cette recherche de flexibilié en proposant d'exploiter la méthode de construction des relations de surclassement effectuée par `ELECTRE III` pour les trois problématiques. Ainsi, les relations de surclassement peuvent être établies, soit entre les hypothèses concurrentes pour former une structure de préférences qui sera ensuite interprétée pour effectuer une sélection ou un classement, soit entre les hypothèses et les limites d'acceptabilité des classes considérées dans une problématique de tri.

Notre objectif étant de disposer d'un ensemble d'outils décisionnels intégrables dans une chaîne de traitement comme `TiLT`, nous verrons dans la section suivante que cette démarche de généralisation de la comparaison des hypothèses autour des relations de surclassement a simplifié la conception d'un module de contrôle opérationnel.

3.3.3 Adaptation et simplification d'aspects techniques

Outre ces particularités liées au contexte décisionnel et à notre recherche d'une méthodologie générique, nous avons procédé à des modifications ou des simplifications d'ordre technique par rapport aux méthodes originales. Nos travaux ont pour premier objectif de montrer l'apport des méthodes d'AMCD pour la résolution de cas d'indétermination, nous n'avons donc pas pris en compte certains aspects techniques liés à l'utilisation des méthodes `ELECTRE` dans des contextes particuliers.

Les modifications énumérées ci-dessous ont donc pour objectif de proposer une méthodologie qui soit plus facilement interprétable par un décideur et de simplifier la tâche de mise en place d'une stratégie de contrôle.

[8] Ce contexte étant défini par la nature des critères et des informations préférentielles qui leur sont associées.

Utilisation d'une méthode commune de comparaison

Bien que les différentes méthodes `ELECTRE` aient été développées pour agréger un certain type de critères et pour répondre à une problématique précise, nous proposons d'exploiter une méthode commune de comparaison des hypothèses pour différents types de critères et pour les trois problématiques envisagées. En effet, bien qu'ayant été principalement développée pour l'agrégation de *pseudo-critères*, la construction des relations de surclassement qui forme la base de la méthode `ELECTRE III` peut également être exploitée pour des *vrai-critères* et des *quasi-critères*. De même, la méthode d'interprétation des relations de surclassement établies par `ELECTRE III` a été définie pour répondre à une problématique de classement. L'ensemble de ces relations peut cependant être également interprété pour identifier un sous-ensemble d'hypothèses à privilégier et donc répondre à une problématique de sélection.

Cette modification majeure dans l'utilisation des méthodes `ELECTRE`s ayant été décrite plus en détails précédemment, nous renvoyons le lecteur à la section 3.3.2.

Stratégie de tri

Nous avons vu dans la section 3.2.3 que l'affectation des hypothèses dans les différentes classes définies *a priori* pouvait suivre une stratégie pessimiste (de la classe n vers la classe 1) ou optimiste (de la classe 1 à la classe n).

Initialement, la mise en place d'une procédure de tri à l'aide de la méthode `ELECTRE TRI` s'appuie sur l'application de ces deux stratégies. Comme lors de la phase de construction d'un pré-ordre partiel à partir de deux pré-ordres totaux pour les problématiques de classement, l'application de ces deux stratégies antagonistes permet de mettre en évidence des situations d'incomparabilité entre des hypothèses et les vecteurs d'acceptabilité des classes. Dans un premier temps, nous avons décidé d'exploiter directement les résultats proposés par l'application d'une stratégie pessimiste. Cependant, si les résultats obtenus lors de l'application de notre approche décisionnelle de contrôle relèvent des incohérences, nous pourrons à faible coût compléter la procédure de tri pour introduire la prise en compte d'hypothèses incomparables.

Indépendance des informations préférentielles vis-à-vis des performances et simplification des paramètres techniques

L'une des particularités des approches par surclassement en AMCD est d'intégrer des informations préférentielles lors de la comparaison des hypothèses concurrentes. Ces informations préférentielles sont associées aux différents critères pour prendre en compte l'imprécision et l'incertitude liées à leurs jugements. Ces différents paramètres : poids, seuils de préférence, d'indifférence et veto, peuvent varier en fonction de la performance atteinte par les hypothèses sur ce critère (par exemple $p_j(g_j(h_i))$. Dans un premier temps, nous ne prenons en compte que des informations préférentielles fixes (par exemple p_j) et indépendantes des performances obtenues par les hypothèses sur le critère j.

De même, nous avons introduit dans la section 3.2.3 un paramètre technique appelé seuil de coupe (α), déterminant la limite à partir de laquelle un indice de crédibilité est considéré comme suffisamment élevé pour que la relation de surclassement soit établie. Dans la mesure où il apparaît relativement peu intuitif de déterminer ce seuil sans connaissance préalable du comportement des méthodes par surclassement, nous avons défini une valeur par défaut à ce

paramètre qui pourra par la suite être modifiée.

De plus, lors de la présentation de la méthode de construction d'un classement (préordre partiel) à partir de la structure de préférences établies par `ELECTRE III` (section 3.2.3), nous avons introduit un paramètre technique désigné seuil de discrimination ($s(\lambda)$), sur lequel s'appuie la détermination des distillats. Ce seuil qui apparaît comme un paramètre très technique et difficile à interpréter a été fixé en reprenant des valeurs communément établies ($\alpha = 0,30$ et $\beta = -0,15$) [Martin and Legret, 2005] et [Roy and Bouyssou, 1993] (page 418).

Simplification de la méthode d'interprétation des structures de préférences en vue d'une sélection

Nous avons vu dans la section 3.2.3, que l'interprétation d'une structure de préférences en vue d'une sélection pouvait nécessiter une étude supplémentaire du noyau identifié (circuits composés de plusieurs sommets). Cette étude supplémentaire permet d'identifier dans un circuit une hypothèse qui possède suffisamment d'arguments pour être privilégiée par rapport aux autres. L'objectif de l'informatisation et de l'automatisation de ces méthodes pour une tâche de contrôle est d'obtenir un résultat directement exploitable, c'est-à-dire de poursuivre la comparaison des hypothèses par une prise de décision.

Dans le cas où le sous-ensemble d'hypothèses résultant de l'application de la procédure de sélection contient plus d'un élément, nous proposons l'alternative suivante. Si l'objectif de la stratégie de contrôle est de réduire la propagation d'hypothèses erronées, nous renvoyons tout le sous-ensemble d'hypothèses jugées comme plus pertinentes. Les différentes hypothèses de ce noyau sont donc considérées comme indifféremment préférables entre elles.

Si le contexte d'application de la stratégie de contrôle nécessite la sélection d'une seule hypothèse, nous calculons la qualification de chacune des hypothèses du noyau et nous renvoyons celle ayant la qualification maximale. En cas de qualifications équivalentes nous renvoyons simplement une des hypothèses (la première de la structure).

Les traces d'application de l'opérateur décisionnel par surclassement contiennent des informations adressées à l'expert sur la stratégie finale de sélection effectuée, afin évidemment de disposer d'une explication du résultat de la décision.

Conclusion de chapitre : d'une démarche méthodologique à une démarche implantatoire

Dans ce chapitre, nous avons présenté l'AMCD et plus particulièrement les approches par surclassement. Nous nous sommes inspiré de ces méthodes pour concevoir et développer un module de contrôle des cas d'indétermination où interviennent plusieurs critères. Nous allons ainsi voir dans le chapitre suivant comment cette implémentation a été réalisée et comment les choix méthodologiques et techniques ont permis l'intégration d'un système de contrôle décisionnel dans la chaîne de traitement `TiLT`, qui garantit son applicabilité aux différents cas d'indétermination soulevés lors de l'analyse linguistique d'un texte.

Par rapport à la définition du contrôle proposée par [Bachimont, 1992], reprise dans la section 2.1.1, on constate qu'en s'appuyant sur une modélisation des préférences de l'expert, les approches par surclassement proposent un cadre méthodologique répondant à la deuxième interrogation posée, à savoir : "Comment utiliser les connaissances de contrôle (i.e. les critères) ?". Ainsi, à

l'aide d'un ensemble de paramètres préférentiels, l'expert peut déterminer comment doivent être utilisés les critères de comparaison disponible. Nous allons voir dans le chapitre suivant qu'une modélisation rigoureuse de notre méthodologie de contrôle nous permet d'obtenir un système déclaratif, où les instructions de manipulation des connaissances de contrôle sont totalement externalisées du système de contrôle.

4

Implémentation d'un système décisionnel de contrôle

Sommaire

4.1 Vers une architecture décisionnelle de contrôle 69
4.1.1 Confronter les besoins de contrôle au processus d'analyse 69
4.1.2 Architecture logicielle de `TiLT` : abstraire pour généraliser 73
4.1.3 Modélisation du module décisionnel 73
4.1.4 Externaliser pour adapter 77
4.1.5 Proposition d'adaptation de l'architecture de traitement en vue d'une externalisation complète des aspects décisionnels 78
4.2 Processus de contrôle 80
4.2.1 Construction du contexte décisionnel 80
4.2.2 Construction des relations de surclassement en vue d'un classement ou d'une sélection 82
4.2.3 Construction des relations de surclassement en vue d'un tri 85
4.2.4 Quelques fonctionnalités supplémentaires liées à la manipulation des critères 85
4.3 Suggérer un modèle de préférences à partir d'un corpus d'hypothèses de références 88
4.3.1 Regard statistique sur le modèle de préférences 88
4.3.2 L'annotation comme l'expression des préférences d'un expert (décideur) 89
4.3.3 Quelques heuristiques pour suggérer un modèle de préférences à partir d'un corpus de référence 92

La présentation du contexte de notre étude nous a conduit à introduire un phénomène problématique qui semble à la fois récurrent et induit par l'automatisation du processus d'analyse linguistique de textes. En effet, bien que la génération et la propagation d'interprétations erronées aient été observées sur un exemple précis de chaîne de traitement, `TiLT`, ce phénomène apparaît comme une limite récurrente aux différents processus d'analyse envisagés et ceci, quelle que soit la tâche ou le niveau d'analyse concerné.

Le contrôle des différents processus d'analyse nécessite la plupart du temps de prendre en compte des sources de jugement hétérogènes et souvent contradictoires, qui une fois combinées, permettraient d'avoir une meilleure évaluation de la pertinence relative des hypothèses concurrentes.

Ce constat nous a ensuite amené à envisager un rapprochement avec les méthodes issues de l'AMCD, dont le principal objectif est d'assister un expert (décideur) lors de la mise en place d'une stratégie décisionnelle exploitant le plus finement possible l'ensemble des sources de jugement disponibles (critères). En effet, à travers l'usage de paramètres décisionnels introduisant des notions d'imprécision (seuils de préférence et d'indifférence), et d'incomparabilité (seuil veto), nous envisageons une exploitation plus réaliste des critères de comparaison disponibles et de ce fait, l'obtention de recommandations de décisions de contrôle plus pertinentes.

Après nous être intéressé aux aspects méthodologiques de notre approche de contrôle des "points de décision", nous abordons au cours de ce chapitre le travail de conception et d'implémentation de cette proposition. Ainsi, après une démarche prospective et méthodologique, nos efforts se sont focalisés sur la modélisation et le développement d'un système décisionnel de contrôle, donnant lieu à la réalisation d'un module (nommé `Beslissing`[9]) dédié à cette tâche de contrôle et qui a été intégré dans la chaîne de traitement `TiLT`. Nous verrons également que l'intégration de ce module à entraîné une adaptation de l'architecture de traitement de `TiLT` et que ce travail nous a conduit à la proposition d'un système d'analyse offrant une cohabitation optimale des processus de traitement et de contrôle. Chercher à rendre opérationnelle une proposition de méthodologie de contrôle nous permet au moins d'attester de sa faisabilité. Les recherches menées dans le cadre de l'obtention d'un doctorat conduisent fréquemment à l'obtention d'un prototype logiciel expérimental. Cependant, en considérant le contexte industriel dans lequel s'inscrivent nos travaux, il nous est apparu important d'accorder une attention particulière à la conception et à l'implémentation de cette méthodologie, en cherchant notamment à disposer à terme d'une solution logicielle opérationnelle, fonctionnelle et conforme aux exigences d'un tel contexte industriel.

Nous concluons ce chapitre par une perspective venant compléter la méthodologie proposée. Nous verrons qu'il a été fréquemment envisagé d'inférer automatiquement une stratégie de contrôle à partir d'un ensemble d'exemples de décisions émises par un expert. Après avoir constaté que les corpus de référence, couramment utilisés en TALN constituent de tels exemples de décisions, nous proposons quelques pistes basées sur des observations statistiques, pour aider un expert lors de la mise en place d'une stratégie de contrôle en lui suggérant des valeurs possibles pour les différents paramètres décisionnels qui accompagnent les méthodes par surclassement.

[9] Beslissing signifie décision en flamand.

4.1 Vers une architecture décisionnelle de contrôle

La mise en place d'une procédure de contrôle dans un système de TALN existant nécessite une réflexion sur différents aspects détaillés. Ainsi après avoir défini les objectifs, les connaissances de contrôle, la façon de les exploiter, il se pose la question de l'architecture qui permettra de compléter le processus initial d'analyse par des fonctionnalités de contrôle. Comme le souligne [Bachimont, 1992], il est impossible de prétendre modéliser une architecture générique pour tous les systèmes de résolution de problème d'IA. [Bachimont, 1992] avec le système `ABACAB` et [Erman *et al.*, 1980] pour le système `HEARSAY II` se sont notamment intéressés à la modélisation des éléments de contrôle dans le cadre dune architecture distribuée de tableau noir.

Bien que `TiLT` repose également sur l'application de modules de traitement spécialisées, les communications entre ces entités sont beaucoup plus simples dans la mesure où elles reposent sur des interdépendances quasiment unidirectionnelles et séquentielles. Nous allons constater que l'architecture de contrôle proposée pour `TiLT` et plus généralement pour ce genre de systèmes modulaires et séquentiels reprend tout de même les nombreux avantages des systèmes distribués, à savoir un contrôle centralisé de l'ensemble des hypothèses construites pour la résolution d'un problème.

Ainsi, lorsque nous utilisons le terme de générique pour qualifier notre système de contrôle, nous désignons évidemment sa capacité à être appliqué sur n'importe quel cas d'indétermination soulevé par `TiLT` et non pas sur n'importe quel problème de contrôle en IA. Nous verrons également que ce qui nous différencie des principaux travaux autour du contrôle est le fait de s'appuyer sur le paradigme de la programmation orientée objets pour atteindre ce degré de généricité et non pas sur le paradigme du multi-agents qui soulève le délicat problème de la communication entre agents.

4.1.1 Confronter les besoins de contrôle au processus d'analyse

Lors de la présentation de la stratégie d'analyse ainsi que de l'architecture logicielle de la chaîne de `TiLT` (section 1.2.2), nous nous sommes focalisé sur les aspects liés à la génération des interprétations linguistiques (hypothèses). La notion de contrôle, que nous avons définie (section 2.1) comme une démarche d'évaluation de la pertinence des hypothèses générées, est également présente dans l'architecture initiale de `TiLT`.

En effet, outre le rôle de contrôle "linguistique" que constitue l'application successive de différents modules d'analyse, de nombreux critères de comparaison ont été envisagés et intégrés dans le processus d'analyse. Ces informations distinctives, qui sont introduites sous forme de traits dans le lexique, de poids hiérarchisant l'application des règles grammaticales, de scores calculés par des méthodes *ad-hoc* d'évaluation, de marques de validité vis-à-vis de propriétés, de préférences empiriques, etc., sont cependant difficilement utilisables et surtout réutilisables. En effet, ces données matérialisées par des variables de classes ou par des paramètres propagés lors des appels aux procédures de traitement sont la plupart du temps inutilisables en dehors du module d'analyse en question. Il apparaît cependant intéressant de connaître la raison de la propagation d'une hypothèse tout au long du processus d'analyse. Pour illustrer cette nécessité, nous pouvons envisager le cas où l'on souhaite identifier les raisons du choix d'une hypothèse d'analyse syntaxique parmi l'ensemble des hypothèses générées. Il est alors indispensable d'avoir à la fois une vision des stratégies de contrôle effectuées lors de l'analyse syntaxique, mais également de connaître les décisions prises précédemment, notamment lors d'une sélection, d'un tri ou d'un

classement des terminaux (unités lexicales regroupées selon leur catégorie morpho-syntaxique) soumis à l'analyse syntaxique.

Ainsi, l'une des principales fonctionnalités attendues de notre système décisionnel est de maintenir une vision complète de l'ensemble des décisions émises lors du processus d'analyse. Cette traçabilité des composants décisionnels repose sur une centralisation de l'ensemble des informations utilisées et générées par notre module décisionnel.

Cette nécessité de centraliser les informations décisionnelles utilisées par un système a déjà été mise en avant dans d'autres contexte que le nôtre. En effet, souvent décrits sous le terme de `Decision support systems` ou `data warehouses for decision support` [Shim *et al.*, 2002], ces composants logiciels sont dédiés au stockage des éléments décisionnels, ainsi qu'à la centralisation de l'ensemble des procédures décisionnelles. Bien que les travaux dans ce domaine soient principalement orientés autour d'aspects économiques, les notions de centralisation et d'accès à ces connaissances que nous souhaitons déployer et intégrer dans la chaîne de traitement `TiLT` répondent bien aux mêmes objectifs que ces entrepôts de données décisionnelles.

La création d'un module de contrôle décisionnel doit également être à l'origine d'une simplification de la création et de la manipulation des données et procédures décisionnelles. La démarche initiale de contrôle consistait à définir des critères de comparaison, à les intégrer dans la description des objets linguistiques, puis à l'écriture du code réalisant l'interprétation de ces critères et prenant la décision de contrôle. Ainsi, à de multiples endroits du code de `TiLT` apparaissent des méthodes visant à ranger, à sélectionner ou bien à trier les hypothèses générées. Cette démarche a évidemment pour conséquence d'alourdir le code, de rompre la réutilisabilité des données et des méthodes et introduit également le risque qu'un des multiples programmeurs intervenant dans le développement de `TiLT` introduise d'éventuels "bugs".

La simplification et la généralisation de la démarche décisionnelle repose évidemment sur l'identification des aspects communs à toutes les procédures de contrôle, ce que nous avons mis en évidence lors de la section précédente, mais également sur l'identification du caractère spécifique que les différents cas de contrôle peuvent entraîner. Cette opposition générique - spécifique nous a conduit à adopter une position visant à externaliser l'ensemble des paramètres permettant d'adapter et de spécialiser une stratégie décisionnelle à un contexte de contrôle. Cette démarche rejoint ainsi la conception fonctionnelle de la chaîne de traitement `TiLT`, où l'ensemble des paramètres liés à la stratégie d'analyse sont regroupés dans des fichiers de configuration.

Nous venons d'esquisser les fonctionnalités apportées par notre système, que nous pouvons résumer autour des notions suivantes :

- généralisation de la notion de critère pour identifier chaque information distinctive ;
- simplification de l'association de critères à des hypothèses d'analyse ;
- généralisation des méthodes de comparaison des hypothèses concurrentes (sélection, classement et tri autour des relations de surclassement) ;
- centralisation de l'ensemble des éléments décisionnels (données et méthodes) ;
- traçabilité et réutilisabilité des actes décisionnels ;
- paramétrabilité et extensibilité des procédures de contrôle à travers l'externalisation des paramètres décisionnels.

Afin de définir les cas d'utilisation de notre module de contrôle décisionnel, il convient de compléter la liste des fonctionnalités attendues par une description des différents acteurs pouvant

intervenir sur ce système. Nous en considérons trois : un humain et deux logiciels. En effet, les choix stratégiques et méthodologiques que nous avons mis en avant dans les deux chapitres précédents sont principalement orientés autour de la place à accorder au décideur. Nous rappelons que le rôle de décideur dans notre système de contrôle décisionnel est joué par le linguiste ou informaticien, qui en tant qu'expert apporte ses connaissances et intuitions sur le cas de contrôle à traiter.

Les deux acteurs logiciels en interaction avec notre module de contrôle décisionnel sont les modules de traitement générant et manipulant des hypothèses linguistiques et les hypothèses elles-mêmes. En effet, les modules de traitement sont à l'origine de l'instanciation des différentes hypothèses concurrentes, mais comportent également les méthodes permettant de les propager aux modules suivants. C'est donc à partir de ces modules de traitement que l'invocation d'une stratégie de contrôle est émise, afin d'évaluer la pertinence des hypothèses générées. Les hypothèses linguistiques correspondent donc aux éléments sur lesquels porte l'action de contrôle et sur lesquels seront affectées des performances de critères. Bien que les hypothèses linguistiques n'apparaissent pas intuitivement comme des acteurs du système, ce rôle leur est tout de même conféré dans la mesure où il apparaît intéressant de pouvoir accéder aux éléments décisionnels qui leur sont associés, et ceci en tous points du système. Nous considérons ainsi les hypothèses linguistiques comme des acteurs du module de contrôle, pouvant à la fois accéder aux critères qui les qualifient mais également modifier les performances atteintes sur ces critères.

Le diagramme de cas d'utilisation suivant, figure 4.1, illustre l'ensemble des fonctionnalités proposées par notre module de contrôle à ces différents acteurs.

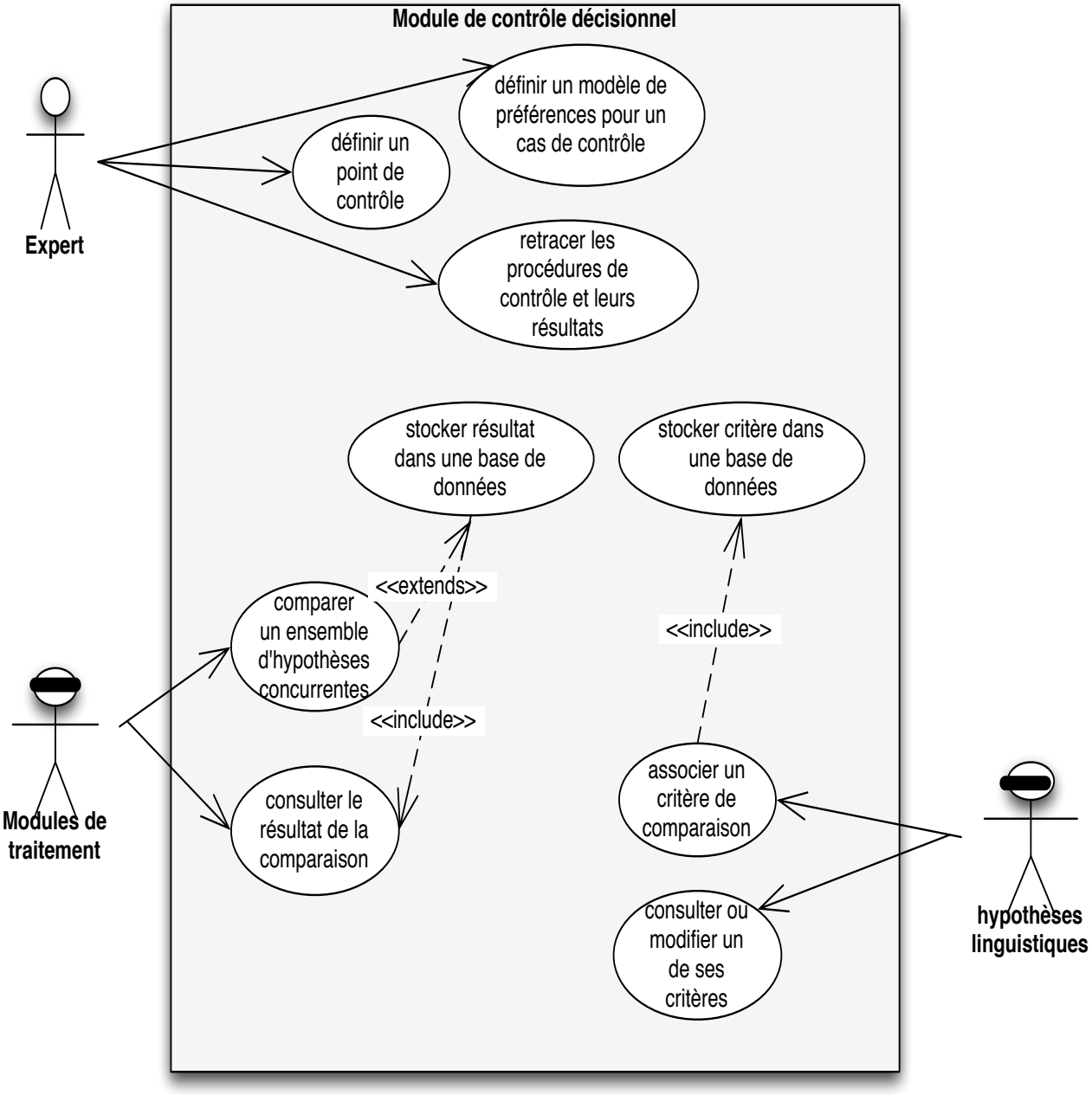

FIG. 4.1 – Cas d'utilisation du module de contrôle décisionnel par les différents acteurs

4.1.2 Architecture logicielle de TiLT : abstraire pour généraliser

Le diagramme des cas d'utilisation présenté sur la figure 4.1 nous a permis d'illustrer à partir d'une vue extérieure l'ensemble des fonctionnalités attendues par les différents acteurs du système. Le processus de modélisation statique du système de contrôle passe désormais par l'identification et la description des éléments (classes) qui le composent. Mais avant de poursuivre la modélisation de notre système de contrôle, il nous faut nous intéresser un peu plus à la structure logicielle de TiLT.

En effet, la difficulté majeure de cette étape de modélisation était de trouver des points d'ancrage dans l'architecture de traitement TiLT permettant d'assurer la généricité, c'est-à-dire l'accessibilité permanente, de notre module de contrôle décisionnel. Intégrer un nouveau développement dans une architecture existante constitue sans doute l'une des tâches les plus compliquées en modélisation et conception des systèmes d'information. Cette tâche est rendue encore plus délicate lorsqu'elle concerne un logiciel comme TiLT, dont l'architecture logicielle résulte de développements effectués par de multiples personnes sur une période de plus de 10 ans.

Dans le diagramme des cas d'utilisation (figure 4.1), nous avons introduit deux acteurs logiciels : les modules de traitement et les hypothèses linguistiques. Cependant, ces éléments ne sont qu'implicitement présents dans l'architecture actuelle de TiLT. En effet, les modules de traitement correspondent aux classes principales appelées lorsqu'une étape précise d'analyse doit être effectuée. Il s'agit par exemple du segmenteur, de l'analyseur lexical, de l'analyseur syntaxique de surface, de l'analyseur syntaxique en dépendances, etc. Quant aux hypothèses linguistiques, il s'agit de l'ensemble des objets générés par ces différents modules. Ce sont ces objets linguistiques qui contribuent à la construction progressive de l'interprétation finale de l'énoncé analysé. Il s'agit par exemple des segments, des unités lexicales, des tronçons (chunks), des arbres de dépendances, etc. ou même d'éléments plus précis tels que des traits ou des prédicats.

Ainsi, dans une démarche d'abstraction et de généralisation de ces notions, nous avons défini ces acteurs logiciels comme des classes de base du système TiLT. Sans remettre en cause le fonctionnement du système actuel, nous avons ajouté simplement une couche d'abstraction permettant de désigner sous des notions communes les principaux éléments de la chaîne de traitement, à savoir les modules de traitement en tant que `TiLTGestionnaire` et les hypothèses linguistiques générées de différentes natures en tant que `TiLTDataObject`.

Cette démarche préalable de généralisation nous a conduit au modèle illustré par la figure 4.2.

Maintenant que nous disposons des acteurs logiciels qui constituent les points d'ancrage de notre système dans la chaîne de traitement TiLT, nous pouvons débuter la modélisation statique de l'architecture du module de contrôle décisionnel.

4.1.3 Modélisation du module décisionnel

La démarche de conception du modèle statique de notre module décisionnel passe désormais par l'identification des différents éléments (classes) nécessaires à la mise en place des fonctionnalités présentées précédemment. Nous avons mis en avant l'intérêt d'une normalisation et d'une centralisation de tous les éléments décisionnels disponibles lors du processus d'analyse et notamment des critères de comparaison. L'élément central du module de contrôle `Beslissing`,

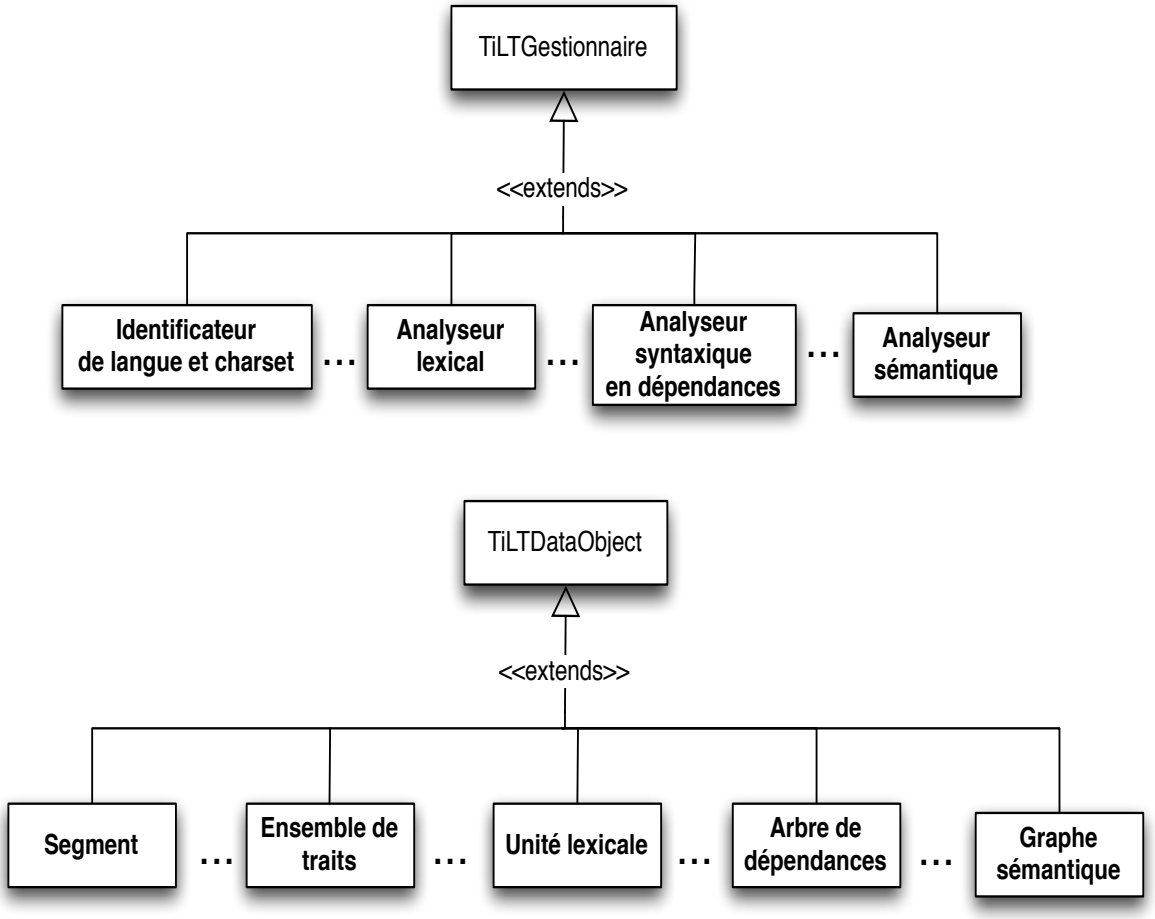

FIG. 4.2 – Identification des acteurs logiciels par abstraction des notions de module de traitement et d'hypothèse linguistique

que nous avons désigné sous le terme de `GestionnaireBeslissing`, constitue ainsi une interface entre les acteurs logiciels (modules d'analyse et hypothèses) et notre système. Ce gestionnaire est notamment consulté pour :

- créer et associer des critères de comparaison aux hypothèses linguistiques ;
- accéder et modifier les performances atteintes par les hypothèses sur ces critères ;
- définir un ensemble d'hypothèses à comparer et appliquer un opérateur décisionnel sur cet ensemble ;
- centraliser / stocker les critères disponibles et les résultats des étapes de contrôle (classement, sélection, tri) ;
- procéder à la desctruction des objets décisionnels instanciés.[10]

Le `GestionnaireBeslissing` est donc composé à la fois de méthodes réalisant l'interfaçage avec les autres éléments du système, mais également de structures de stockage pour les différents critères instanciés, les structures de préférences établies et les résultats de leur interprétation (pré-ordres, sous-ensembles ou tris). Afin de garantir l'unicité et l'accessibilité complète de ce gestionnaire, nous avons utilisé le patron de conception (design pattern) `Singleton` [Gamma *et al.*, 1980].

En étant accessible pour tous les modules de traitement (instance unique statique), le `GestionnaireBeslissing` constitue alors le point d'entrée du module de contrôle, notamment pour l'instanciation des critères (classe `Critere` et ses sous-classes `CritereNumerique`, `CritereBinaire` et `CritereSymbolique`) ainsi que leur accès dans la structure de stockage dédiée à ces éléments décisionnels.

Lors de la description des fonctionnalités accessibles à partir du `GestionnaireBeslissing`, nous avons introduit la nécessité de regrouper les hypothèses concurrentes devant être évaluées lors d'une phase de contrôle. Ce regroupement effectué au sein d'une `StructureDeComparaison` réalise dans un premier temps une vérification que les hypothèses à comparer partagent des critères de comparaison en commun avant de procéder à la création d'`EnsembleDeCriteres` (les critères utilisés étant définis dans un fichier de configuration, voir section 4.1.4), mais également d'effectuer éventuellement un premier filtrage des hypothèses en éliminant des hypothèses qualifiées par une performance trop faible sur un des critères en question. Cette possibilité de filtrer *a priori* les hypothèses introduites dans la structure de comparaison permet de réduire le nombre de relations de surclassement à construire et interpréter (pour les problématiques de sélection et de classement).

L'appel d'une procédure de classement ou de sélection, toujours à partir du `GestionnaireBeslissing`, s'appuie sur une `StructureDeComparaison` pour établir les `RelationDeSurclassement` entre les différentes `HypotheseComparee`, et former ainsi une `StructureDePreferences`. Cette `StructureDePreferences` est donc ensuite interprétée par un `OperateurDeSelection` ou un `OperateurDeClassement` pour établir respectivement un `SousEnsembleDHypotheses-Privilegiees` ou un `PreOrdrePartiel`. L'`OperateurDeTri` exploite également une définition des limites d'acceptabilité des différentes classes considérées afin de trier les hypothèses de la structure de comparaison et retourner un `Tri`. Ces différents résultats sont également stockés par le `GestionnaireBeslissing` pour à la fois pouvoir y accéder par la suite afin de prendre des décisions de propagation ou de filtrage mais également afin d'être en mesure de retracer les décisions de contrôle émises lors du processus d'analyse.

[10] `TiLT` étant développé en `C++`, la gestion de la destruction des objets instanciés doit être explicitement réalisée.

Nous venons de présenter comment les différents acteurs logiciels utilisaient par l'intermédiaire du `GestionnaireBeslissing` les différents éléments décisionnels qui forment une stratégie de contrôle. Nous avons cependant également évoqué à plusieurs reprises l'usage d'une configuration permettant notamment de déterminer les critères de comparaison à prendre en compte et de définir le comportement du processus de comparaison. L'informaticien ou linguiste jouant le rôle d'expert / décideur spécifie le comportement attendu lors d'une phase de contrôle à l'aide des différents paramètres externalisés dans un fichier de configuration. Avant de procéder à une description de ces paramètres et de leur usage, nous pouvons introduire un dernier élément à notre module décisionnel chargé d'effectuer la lecture et l'interprétation d'un fichier de configuration : le `LecteurDeConfiguration`.

La figure 4.3 illustre une vue simplifiée du module décisionnel `Beslissing` avec ses points d'ancrage dans l'architecture de traitement existante de `TiLT` (une version plus détaillée du diagramme de classes est proposé en annexe B.1 page 179.

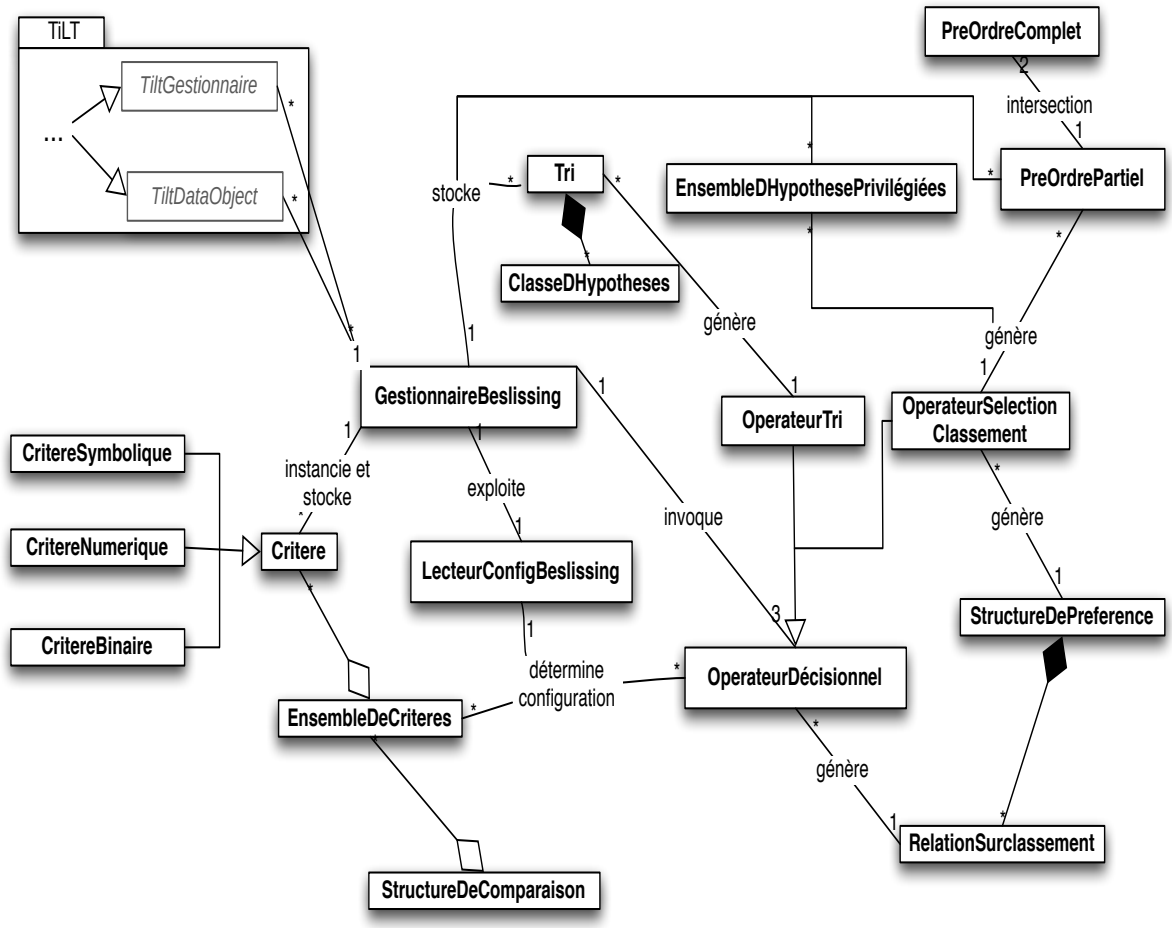

FIG. 4.3 – Modèle du domaine du module de contrôle décisionnel

4.1.4 Externaliser pour adapter

Comme nous l'avons souligné à plusieurs reprises lors de la présentation du modèle statique du module de contrôle `Beslissing`, l'adaptabilité des opérations de contrôle aux particularités des différents cas d'indétermination repose sur l'externalisation de l'ensemble des paramètres décisionnels. Ainsi, le fichier principal de configuration de `TiLT` contient désormais une section complémentaire dédiée aux paramètres décisionnels de contrôle. Cette section illustrée en annexe B.2 page 181 permet notamment :

- de déterminer les critères pouvant être manipulés lors d'une étape de contrôle ;
- les familles/ensembles de critères à construire pour qualifier des hypothèses concurrentes ;
- le modèle de préférences associé aux critères manipulés (poids, seuils de préférence, indifférence et véto, coupe alpha, limites d'acceptabilités pour les problématiques de tri) ;
- les opérations de contrôle : spécifiées par un identifiant, une problématique visée (classement, sélection, tri), un ensemble de critères ainsi que le modèle de préférences à utiliser.

Ainsi, comme l'illustre plus en détail le diagramme de séquences (figure 4.7 page 84), l'application d'un opérateur décisionnel dans un module de traitement nécessite uniquement de disposer d'un regroupement (`StructureDeComparaison`) d'hypothèses concurrentes qualifiées par des critères de comparaison communs, puis d'invoquer l'opérateur décisionnel avec un identifiant d'opérateur défini dans la configuration. Le module se charge ensuite de vérifier que les hypothèses à comparer disposent des critères définis dans la configuration et de construire les relations de surclassement en s'appuyant sur le modèle de préférences également externalisé dans la configuration.

La mise en place d'une stratégie de contrôle par expert peut ainsi résulter d'un processus itératif de modification, d'application et de validation d'une configuration, sans avoir à remettre en cause, et donc recompiler, le code de l'application.

Afin d'éviter des erreurs de format dans le fichier de configuration de `Beslissing`, dont la syntaxe est particulièrement peu intuitive et difficilement lisible, nous avons développé une interface graphique dédiée à la manipulation de ces paramètres décisionnels (annexe C.1). L'usage de cette interface graphique permet à la fois de garantir la validité de la syntaxe du fichier de configuration généré, mais permet également de vérifier la cohérence des paramètres utilisés, comme à titre d'exemple :

- la disponibilité des critères utilisés pour former des ensembles de critères ;
- la cohérence du modèle de préférences : la dimension des vecteurs de poids, seuils de préférence, d'indifférence et de véto est égale au nombre de critères utilisés ;
- la cohérence des seuils : seuils d'indifférence inférieur ou égale au seuil de préférence, lui-même inférieur ou égale au seuil véto ($q_i \leq p_i \leq v_i$) ;
- la déclaration d'opérateurs de tri nécessite de disposer de profils de classes, dont la dimension est égale au nombre de critères utilisés ;
- etc.

Cette démarche déclarative matérialisée par une externalisation des instructions d'utilisation des critères de comparaison disponibles ainsi que la mise en place de fonctionnalité générique de manipulation des éléments décisionnels nous permettent de répondre à la troisième question soulevée par la mise en place d'une stratégie de contrôle (voir section 2.1.1) : "Comment intégrer les connaissances de contrôle dans le système ?" ([Bachimont, 1992] page 145).

En effet, l'instanciation et la manipulation des critères de comparaison repose désormais sur une simple invocation des méthodes dédiées à ces fonctionnalités, méthodes qui sont désormais disponibles en tous points des processus de TALN.

4.1.5 Proposition d'adaptation de l'architecture de traitement en vue d'une externalisation complète des aspects décisionnels

Lors de la présentation du modèle statique du système de contrôle, nous avons insisté sur les différents éléments décisionnels manipulés par les deux acteurs logiciels identifiés, à savoir les modules de traitement et les hypothèses linguistiques. Nous verrons notamment dans la section suivante 4.2 que cette première vision de notre système permet d'intégrer dans le code de la chaîne de `TiLT` des étapes de contrôle. Cependant, le linguiste en tant qu'expert du domaine et donc du cas d'indétermination en question, n'a pas forcément les connaissances techniques requises pour intégrer dans le code (écrit en `C++`) ces phases de contrôle. Nous avons donc envisagé et partiellement développé une adaptation de l'architecture de traitement de `TiLT` permettant d'intégrer des phases de contrôle lors du processus d'analyse sans nécessiter de modification importante (uniquement instancier et associer les critères aux hypothèses) du code de la chaîne de traitement.

Lors de la présentation de la chaîne de traitement `TiLT` (section 1.2.2 page 12), nous avons vu que le comportement du processus d'analyse était déterminé par une stratégie, définissant l'ordre d'application des modules de traitement nécessaires à l'obtention du niveau d'interprétation linguistique souhaité. Bien que le modèle proposé précédemment permette l'appel à des opérateurs de contrôle à l'intérieur même d'un module d'analyse, pour comparer par exemple des hypothèses linguistiques intermédiaires, le contrôle a pour objectif principal d'éviter la propagation d'hypothèses linguistiques erronées d'un module à l'autre. Or, la séquence d'appel des modules de traitement étant définie dans une stratégie d'analyse, externalisée dans un fichier de configuration, nous avons envisagé la possibilité d'introduire dans cette stratégie les phases de contrôle. Une stratégie d'analyse déployée pour un contexte applicatif donné définit ainsi la séquence de modules d'analyse à appliquer sur le texte à analyser, mais également les phases de contrôle à appliquer au cours ou en fin d'analyse pour évaluer la pertinence des hypothèses générées.

Ainsi, en s'appuyant sur un fichier de configuration définissant les critères à utiliser et un modèle de préférences (voir section 4.1.4), cette architecture propose un moyen de contrôler la pertinence des hypothèses transmises d'un module d'analyse à un autre. Comme l'illustre la figure 4.4, ce contrôle complètement externalisé nécessite de disposer d'une centralisation et d'une encapsulation des méthodes d'accès aux hypothèses linguistiques générées au cours du processus d'analyse. En effet, c'est lorsqu'un module d'analyse récupère les hypothèses linguistiques générées par l'application des modules précédents que la stratégie d'analyse doit être consultée pour voir si une étape de contrôle a été spécifiée et quelle configuration (choix des critères et modèle de préférence) doit être utilisée. Cette démarche de contrôle n'a pu être envisagée que dans la mesure où quasiment la totalité des différents types d'hypothèses générées (des segments aux chunks) sont stockées et centralisées dans une structure commune correspondant à un treillis d'analyses. La généricité de cette `API` d'accès contrôlé se limite donc aux hypothèses stockées dans le treillis d'analyse et devient spécifique pour les autres types d'hypothèses (arbres syntaxiques de dépendance et graphes sémantiques).

Ainsi, ce contrôle lors de l'accès aux différents types hypothèses linguistiques est transparent pour les modules de traitement, qui appliquent leur processus complémentaire d'analyse uniquement sur les hypothèses non filtrées et éventuellement dans l'ordre établi par un opérateur de classemenent. L'encapsulation des méthodes d'accès aux différentes structures de stockage des hypothèses générées par les différents modules de traitement nous rapproche des architectures de traitement centralisant l'ensemble des analyses comme les tableaux noirs. En proposant un tri, une sélection ou un classement des résultats retournés par ces méthodes d'accès aux hypothèses intermédiaires, nous reproduisons le fonctionnement du contrôleur en charge de veiller à la progression du problème et à la pertinence des hypothèses générées [Bachimont, 1992].

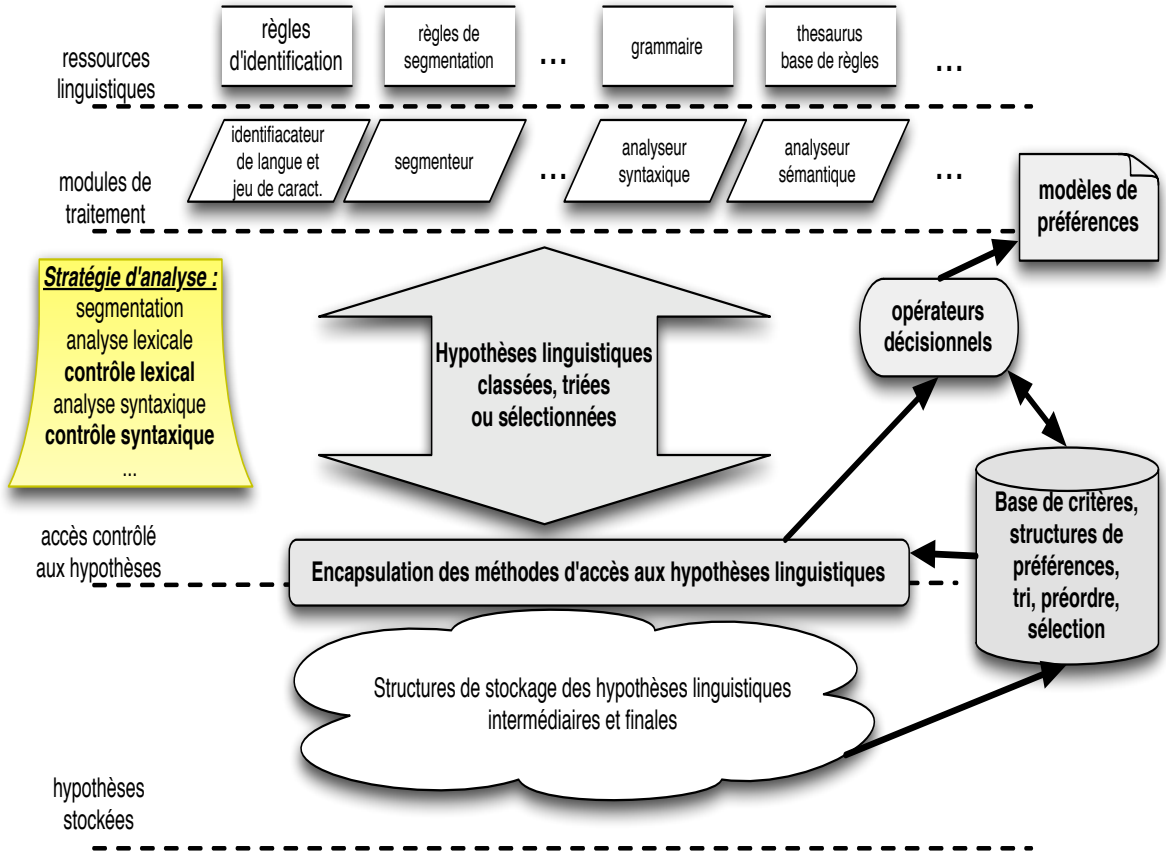

FIG. 4.4 – Vers une architecture de traitement contrôlé

Cette architecture de traitement contrôlée permet ainsi à un expert (linguiste) de déployer des stratégies de contrôle par filtrage ou classement des hypothèses concurrentes, ceci en différents points du processus d'analyse et sans modifier le code de l'application. Cependant, contrairement au module de contrôle `Beslissing` qui a donné lieu à une implémentation opérationnelle et validée dans la chaîne de traitement commerciale de `TiLT`, cette proposition d'architecture a uniquement été développée dans le cadre d'un prototype de recherche. Ces développements sont restés au stade expérimental pour à la fois des raisons de temps mais également parce qu'ils nécessitaient une modification plus importante du code existant de `TiLT`.

4.2 Processus de contrôle

Afin de mieux comprendre l'apport de notre système et notamment la simplification qu'il apporte pour la mise en œuvre d'une stratégie de contrôle, cette section s'appuie sur des diagrammes de séquence qui décrivent en détail les principaux cas d'utilisation introduits précédemment (section 4.1.1 figure 4.1). Dans un premier temps, nous présentons le processus de mise en place d'un contexte décisionnel, qui repose principalement sur l'identification des hypothèses à comparer ainsi que le chargement du modèle de préférences à utiliser. Nous verrons ensuite que cette méthode commune de déclaration du contexte décisionnel précède l'appel à un opérateur décisionnel de classement, de sélection ou de tri. Cette section constitue donc le passage du modèle statique au modèle dynamique de notre système de contrôle.

4.2.1 Construction du contexte décisionnel

Nous avons débuté le chapitre 2.1 par une description de ce que nous nommons un cas d'indétermination. Désigné également sous le terme de point de décision, cet état particulier du processus d'analyse est donc caractérisé par l'existence d'hypothèses concurrentes, associées à un vecteur de performances permettant d'appréhender leur pertinence respective.

Ainsi, la première étape de la mise en place d'une stratégie de contrôle repose sur la définition de ce contexte décisionnel. Cette tâche incombe aux différents modules de traitement qui ont été généralisés en tant que `TiltGestionnaire` (voir section 4.1.2). Sous la notion de `StructureDeComparaison`, un module de traitement (`TiltGestionnaire`) regroupe les hypothèses concurrentes `TiltDataObject` qu'il a générées, en spécifiant également l'identifiant de l'opérateur qui va être appliqué sur la `StructureDeComparaison`. Cet identifiant permet de référencer une configuration d'opération de contrôle définie dans le fichier de configuration du module décisionnel (section 4.1.4 et annexe B.2). La configuration d'une opération de contrôle détermine notamment quels sont les critères à utiliser, mais également le modèle de préférences à exploiter. Ainsi, lorsqu'un module de traitement insère une nouvelle hypothèse dans une structure de comparaison, le système vérifie que les critères sont disponibles sur cette hypothèse, et construit alors le vecteur de performances qui lui est associé. De plus, afin de limiter le nombre d'hypothèses comparées, des limites d'acceptabilité peuvent être définies pour associer à chaque critère utilisé la valeur minimale requise afin d'être accepté dans la structure de comparaison.

Pour chaque étape de contrôle à effectuer une instance de la classe `StructureDeComparaison` est alors créée. Comme le montre la figure 4.5, cet objet contient l'ensemble des données et paramètres nécessaires à l'application d'un opérateur décisionnel.

Il est évident que cette démarche de construction d'un contexte décisionnel reposant sur la définition d'une structure de comparaison n'a d'intérêt que si les hypothèses concurrentes ont été évaluées sur les critères de comparaison en question. L'instanciation des critères de comparaison est également initiée par les différents modules de traitement (`TiltGestionnaireObject`), qui définissent quelles informations distinctives doivent être considérées comme des critères de comparaison, ces informations pouvant être directement accessibles (par exemple : des traits lexicaux, le nombre de relations d'un arbre syntaxique, etc.) ou bien résulter de l'application d'une méthode *ad-hoc* de jugement (par exemple : associer une probabilité calculée sur corpus à une hypothèse, vérifier le respect d'une propriété, etc.). La déclaration d'un critère sur une hypothèse concurrente par un module de traitement donne lieu au stockage de cette information décisionnelle dans la base des critères disponibles. Cette centralisation des critères permet éga-

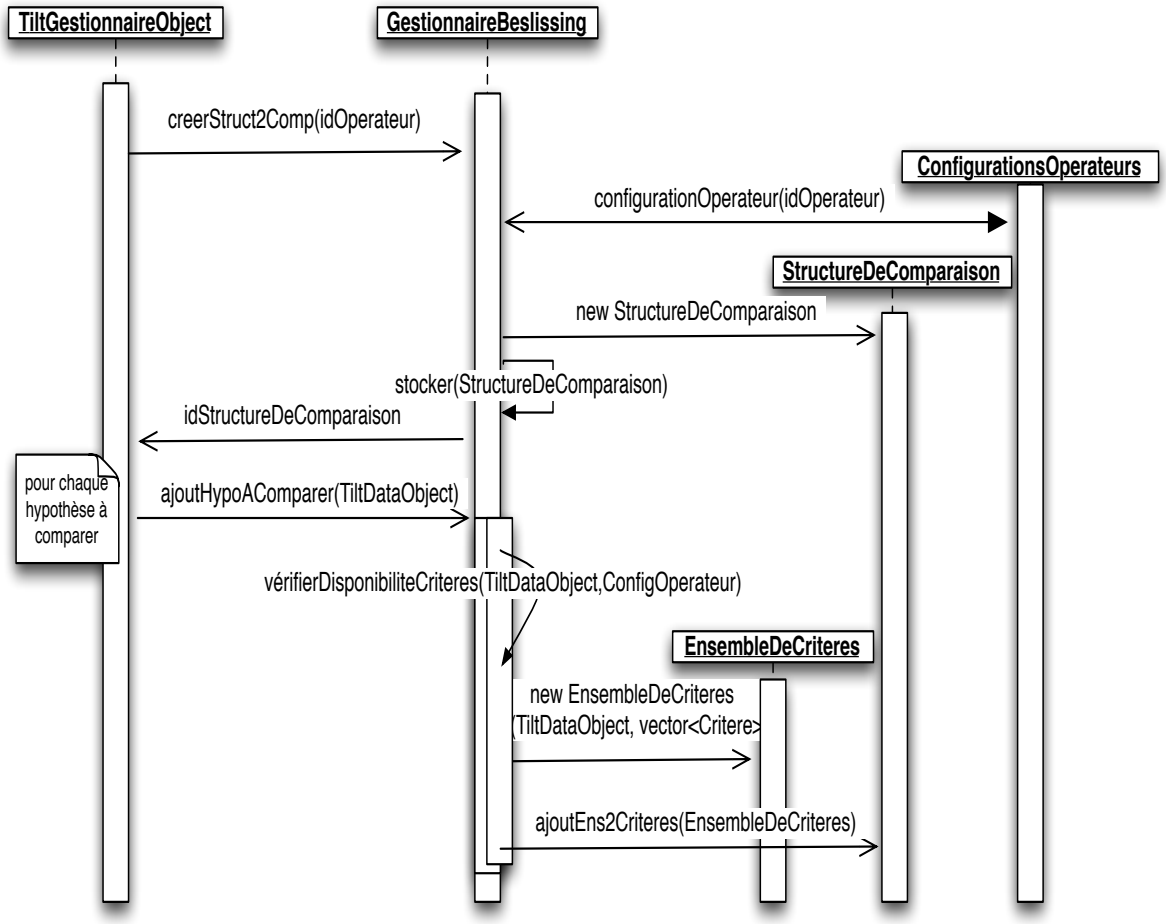

Fig. 4.5 – Diagramme de séquence : Regroupement des hypothèses comparées et leurs critères associées au sein d'une structure de comparaison

lement d'avoir un accès tout au long du processus, que ce soit pour les modifier ou les consulter comme l'illustre la figure 4.6.

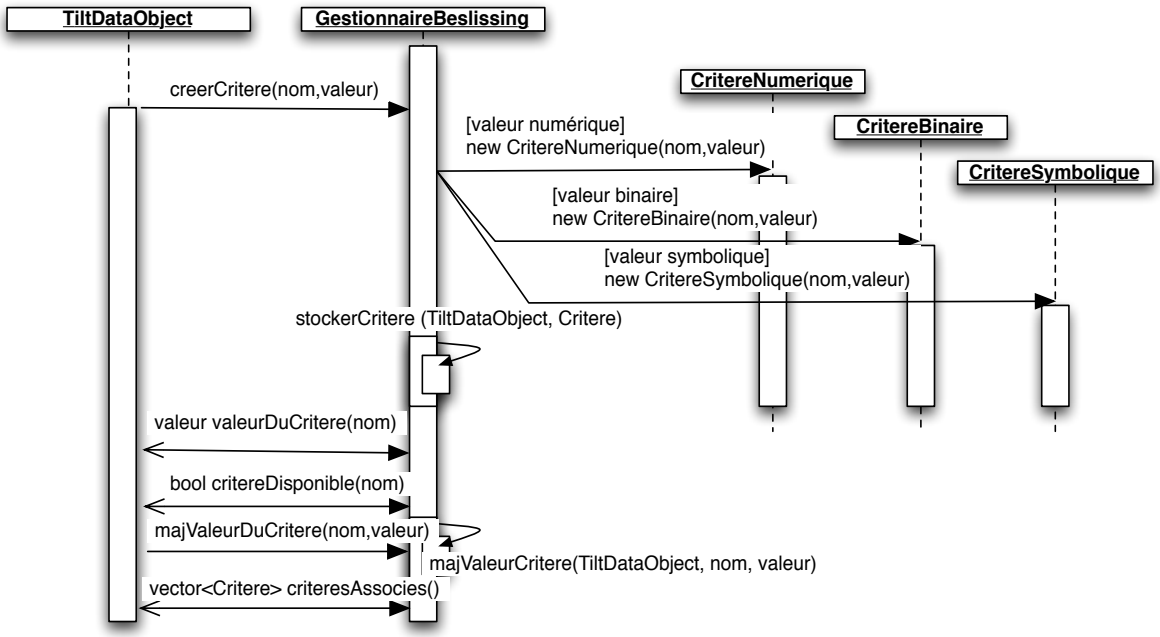

FIG. 4.6 – Diagramme de séquence : Instanciation et stockage de critères associés aux hypothèses linguistiques

4.2.2 Construction des relations de surclassement en vue d'un classement ou d'une sélection

Une fois que la `StructureDeComparaison` est établie, un module de traitement (`TiltGestionnaire`) peut lui appliquer un opérateur décisionnel (`OperateurClassementSelection`). Nous avons introduit précédemment (section 3.2.3) le fait que l'application d'un opérateur de surclassement en vue de répondre à une problématique de classement ou de sélection s'appuyait sur une structure commune, désignée sous le terme de `StructureDePréférences`, dont l'interprétation permettait d'établir une recommandation de classement (`PreordrePartiel`) ou de sélection (`SousEnsembleDHypothesesPreferees`). Une `StructureDePréférences` contient donc l'ensemble des `RelationDeSurclassement` établies entre les différentes hypothèses à comparer référencées dans la `StructureDeComparaison`.

Ainsi, qu'il s'agisse de répondre à une problématique de classement ou de sélection, le processus d'application d'un opérateur de surclassement débute par une étape commune de construction des relations de surclassement entre les hypothèses concurrentes. La distinction entre ces problématiques, et donc les recommandations générées, est effectuée par le module de traitement (`TiltGestionnaire`) qui demande soit l'obtention d'un pré-ordre partiel (`PreOrdrePartiel` : classement avec ex-aequo), soit l'obtention d'un sous-ensemble d'hypothèses privilégiées (`SousEnsembleDHypothesesPreferees`). Comme nous l'avons souligné dans la section dédiée aux mé-

thodes décisionnelles par surclassement (section 3.2.3), la comparaison et donc l'évaluation de la pertinence des hypothèses dépendent à la fois du vecteur de performances associé aux hypothèses, mais également du modèle de préférences qui définit la manière dont les performances de critères doivent être interprétées et agrégées. Afin de rendre générique et adaptable le module de contrôle, nous avons vu que l'ensemble des paramètres décisionnels est externalisé dans un fichier de configuration. Ainsi, lorsqu'un module de traitement invoque la méthode d'application d'un opérateur décisionnel sur une `StructureDeComparaison`, l'identifiant d'opérateur spécifié lors de l'instanciation de la `StructureDeComparaison` est utilisé pour retrouver le modèle de préférences associé à cet identifiant dans la configuration.

Après vérification de la conformité et de la validité du modèle de préférences chargé depuis la configuration, le processus d'application d'un opérateur décisionnel construit les relations de surclassement (`RelationDeSurclassement`) entre les hypothèses pour former une structure de préférences (`StructureDePreferences`). La demande d'un classement ou d'une sélection d'hypothèses à partir d'une `StructureDePreferences` entraîne la construction et le stockage d'un pré-ordre partiel ou d'un sous-ensemble d'hypothèses.

Pour le cas d'une problématique de classement, l'opérateur en charge de répondre à cette tâche suit fidèlement le processus d'interprétation de la structure de préférences définie par la méthode `ELECTRE III` (pour un rappel de la méthode, voir 3.2.3 et annexe A.4). Ce qui signifie que l'obtention du pré-ordre partiel (`PreOrdrePartiel`) correspond à l'intersection de deux pré-ordres complets (`PreOrdreComplet`), l'un ascendant l'autre descendant. Une fois ce pré-ordre partiel médian obtenu, le module de traitement peut accéder aux N meilleures hypothèses en spécifiant la structure de comparaison à l'origine de l'étape de contrôle et ensuite modifier l'ordre de propagation des hypothèses concurrentes pour la suite du processus d'analyse.

En ce qui concerne la problématique de sélection, la demande de récupération d'un `SousEnsembleDHypothesesPreferees` entraîne l'interprétation de la `StructureDePréférences` en utilisant la même démarche que celle effectuée par la méthode `ELECTRE I`, à savoir l'identification du noyau du graphe formé par l'ensemble des relations de surclassement.

La figure 4.7 décrit graphiquement l'enchaînement des invocations de méthodes et des constructions des différents éléments concernés.

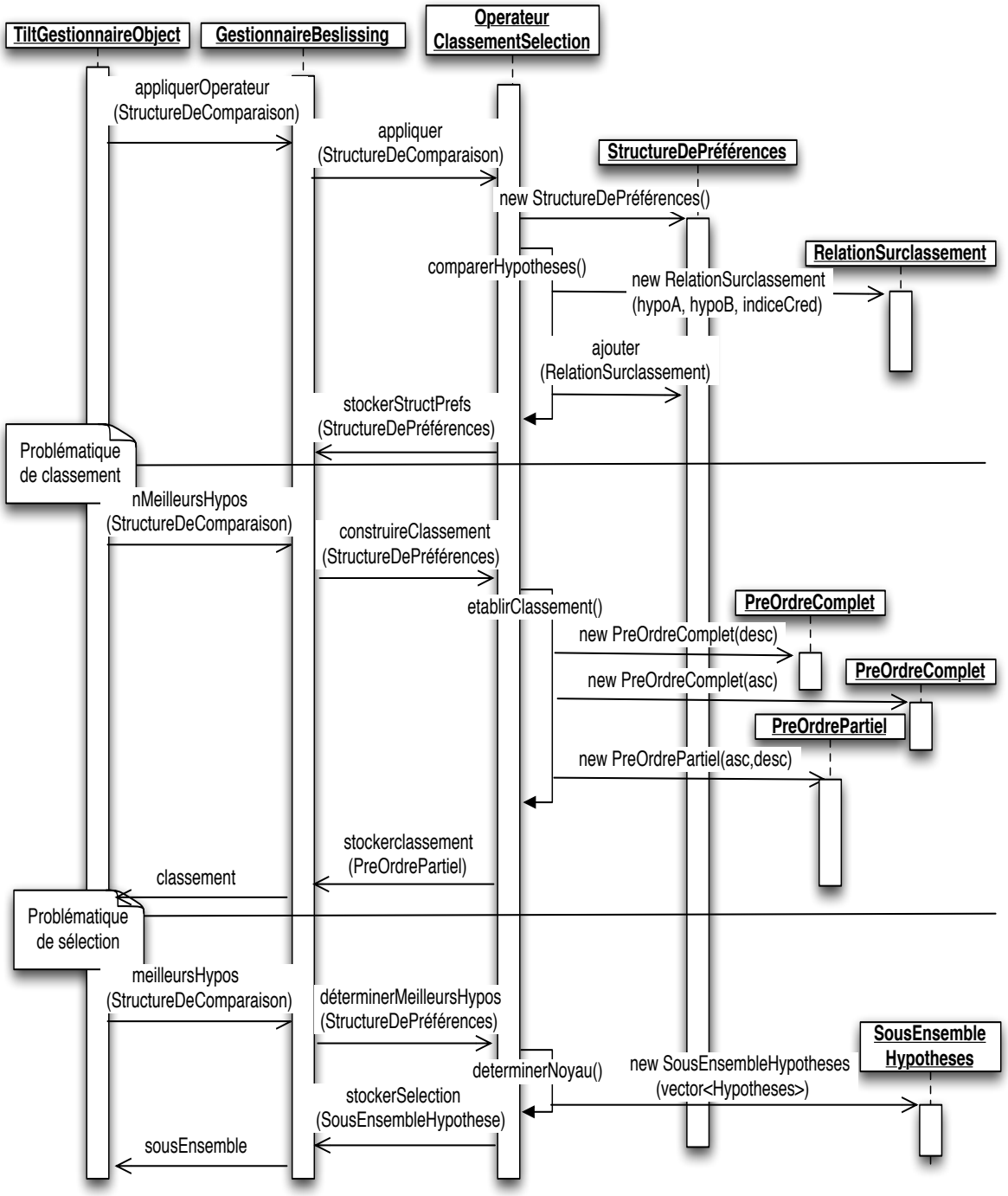

FIG. 4.7 – Diagramme de séquence : processus d'application d'opérateurs de surclassement pour des problématiques de classement ou de sélection

4.2.3 Construction des relations de surclassement en vue d'un tri

L'application d'un opérateur de surclassement en vue de répondre à une problématique de tri est également basée sur cette étape préliminaire de définition du contexte décisionnel (voir section 4.2.1). Cependant, en plus de la nécessité de disposer de l'ensemble des hypothèses concurrentes, de leurs critères associés et du modèle de préférences défini pour l'ensemble de critères concernés, l'application d'un `OperateurDeTri` repose également sur la définition d'un ensemble de limites d'acceptabilité définissant les conditions d'insertion dans les différentes classes définies *a priori*. Ainsi, lorsqu'un module de traitement (`TiltGestionnaire`) invoque la méthode d'application de l'opérateur de tri (`OperateurDeTri`) sur la structure de comparaison (`StructureDeComparaison`) contenant l'ensemble des hypothèses (`TiltDataObject`) à trier, le gestionnaire du module de contrôle vérifie que la configuration associée à la structure de comparaison contient bien la définition des profils d'acceptabilité.

Une fois la conformité de la configuration validée, le processus de tri effectué par l'`OperateurDeTri` reprend fidèlement la méthodologie d'`ELECTRE TRI` telle qu'elle a été présentée lors de la section 3.2.3. Contrairement aux versions d'`ELECTRE` dédiées aux problématiques de classement et de sélection, la construction des relations de surclassement effectuée par `ELECTRE TRI` permet d'obtenir directement une recommandation au problème de tri étudié. En effet, comme nous l'avons vu précédemment, les relations de surclassement sont construites non pas entre les hypothèses concurrentes, mais entre les hypothèses et les profils d'acceptabilité des classes concernées par le tri. La validation d'une relation de surclassement entre une hypothèse et un profil de classe, qui est établie lorsque l'indice de crédibilité calculé est supérieur au seuil de coupe, détermine la classe d'affectation de l'hypothèse.

Ainsi, pour chacune des hypothèses contenues dans la structure de comparaison, l'opérateur de tri tente d'établir une relation de surclassement avec les profils d'acceptabilité des classes, en commençant par la classe la plus haute, c'est-à-dire celle dont les limites d'acceptabilité définies sur chaque critère sont les plus élevées. L'opérateur de tri construit donc progressivement des classes d'hypothèses (`ClassesHypotheses`) qui composent au final le tri (`Tri`). Ce résultat est ensuite stocké par le gestionnaire du module de contrôle (`GestionnaireBeslissing`) pour pouvoir être consulté par les différents modules de traitement (`TiltGestionnaireObject`) et notamment celui qui a initié la procédure de tri. La consultation de ce tri par un module de traitement peut s'effectuer de deux manières : soit en demandant la classe d'affectation d'une hypothèse particulière, soit en demandant la liste des hypothèses qui composent une classe donnée.

Le processus d'application d'un opérateur que nous venons de décrire est illustré par le diagramme de séquences sur la figure 4.8.

4.2.4 Quelques fonctionnalités supplémentaires liées à la manipulation des critères

Nous venons de présenter les procédures d'application des opérateurs de surclassement permettant d'obtenir un classement, une sélection ou un tri à partir d'un ensemble d'hypothèses concurrentes. Outre ces scénarios directement liés à notre objectif de contrôle, c'est-à-dire éviter au plus tôt la propagation d'hypothèses erronées, nous avons développé des fonctionnalités supplémentaires permettant de simplifier la manipulation des différents éléments décisionnels et plus précisément des critères de comparaison.

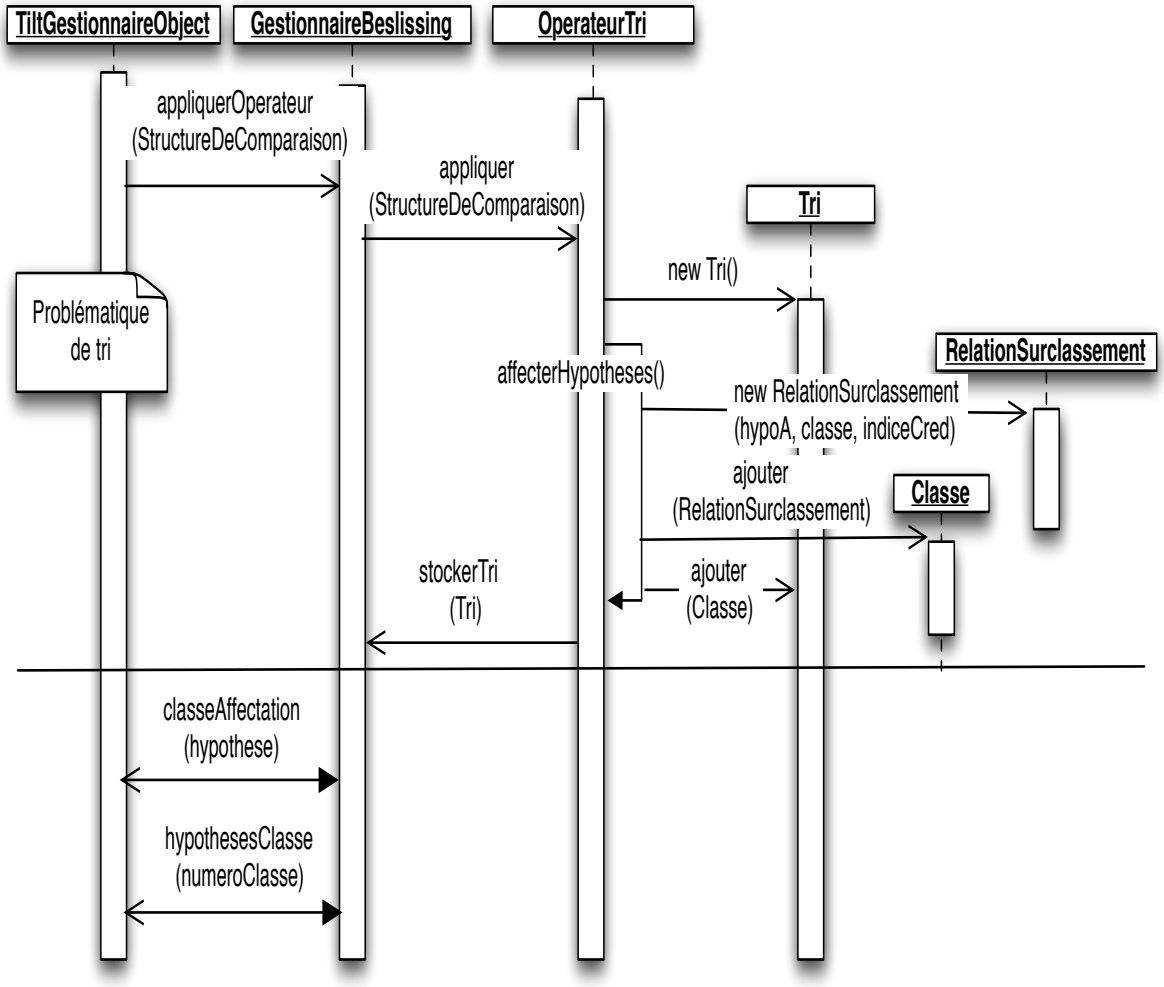

FIG. 4.8 – Diagramme de séquence : processus d'application d'un opérateur de tri par surclassement

À partir d'un cas concret d'application d'une stratégie de contrôle, expérimentation que nous décrirons en détails dans la section 5.3 page 130, nous avons éprouvé la nécessité de développer des fonctionnalités complémentaires afin notamment de pouvoir :

1. dupliquer les critères d'une hypothèse sur une autre (héritage des critères) ;
2. ou synthétiser des ensembles de critères en un ensemble de critères.

Cette première fonctionnalité supplémentaire permet de propager des critères associés à une hypothèse vers une hypothèse dont il est légitime de penser que sa pertinence puisse être évaluée selon ces mêmes performances de critères. Bien que cet exemple soit expliqué avec plus de précision dans le chapitre suivant (section 5.3 page 130), nous avons utilisé cette fonctionnalité de duplication pour propager les critères associés aux unités lexicales résultant de l'accès au lexique, ainsi que de l'application de différentes méthodes de correction, vers les hypothèses qui les factorisent selon leur catégorie morpho-syntaxique, à savoir les terminaux. Ainsi, il est possible de faire hériter à une hypothèse des performances de critères initialement associées à une des hypothèses intermédiaires ayant contribué à sa construction.

Cependant, dans de nombreux cas, la génération d'une hypothèse suite à l'application d'un module de traitement ne s'effectue pas forcément à partir d'une seule hypothèse intermédiaire sous-jacente, mais plusieurs. Par exemple, la construction d'un groupe syntaxique (chunk) s'appuie sur l'interprétation de plusieurs terminaux. Ainsi, si l'on souhaite que ce groupe syntaxique créé puisse hériter des performances de critères définis sur les terminaux qui le composent, il faut nécessairement avoir recours à une procédure d'agrégation des différents ensembles de critères qui qualifient ces terminaux. Nous avons donc développé un opérateur supplémentaire dit d'agrégation permettant de synthétiser ces ensembles de critères, en s'appuyant sur des paramètres définis dans le fichier de configuration du module décisionnel `Beslissing`. Ces paramètres décrivent notamment si l'on conserve la performance minimale, maximale, moyenne obtenue sur chacun des critères à agréger.

L'ajout de ces deux fonctionnalités supplémentaires nous permet à la fois d'exploiter finement la centralisation des éléments décisionnels et de faciliter la manipulation des critères au cours du processus décisionnel, mais permettent également de simplifier l'implémentation des classes désignant les différentes hypothèses linguistiques (`TiltDataObject`). En effet, avant l'intégration de notre système de contrôle générique, la propagation d'informations distinctives entre les différentes hypothèses construites par les modules de traitement nécessitait obligatoirement le passage de toutes ces données en tant que paramètres des constructeurs d'objets et se matérialisait donc par un nombre important de variables de classes. La centralisation de chaque critère établi et la généralisation des objets principaux du système (modules et hypothèses), nous permet de garantir l'accessibilité de ces connaissances décisionnelles tout au long du processus d'analyse et ceci pour tous les modules de traitement définis en tant que `TiltGestionnaireObject`.

Cette section, bien que s'appuyant sur des techniques de modélisation d'ingénierie, nous permet de valider la faisabilité de notre proposition d'une architecture de traitement linguistique contrôlée, où les différentes fonctionnalités permettant d'évaluer la pertinence des hypothèses générées peuvent être utilisées lors de différentes étapes de traitement et pour différentes problématiques.

4.3 Suggérer un modèle de préférences à partir d'un corpus d'hypothèses de références

Afin de répondre aux différentes problématiques de contrôle rencontrées lors du déploiement d'un processus de traitement linguistique, nous avons proposé d'exploiter le paradigme de l'aide multicritère à la décision (chapitre 3) pour notamment disposer d'une méthodologie générique et directement exploitable par un expert du domaine concerné (linguistique ou informaticien). Le principal avantage de notre approche est donc de permettre à un expert d'intégrer sous forme de modèle de préférences ses connaissances et intuitions définissant le comportement du processus de contrôle à mettre en place. Ainsi, contrairement à la plupart des stratégies de contrôle envisagées jusqu'ici, l'utilisation du module de contrôle décisionnel que nous avons développé (section 4.2) n'est pas conditionnée par la disponibilité d'un corpus d'apprentissage suffisamment pertinent et volumineux pour la tâche de contrôle en question.

Cependant, en cas de disponibilité d'un corpus représentatif du cas de contrôle concerné, il semble intéressant d'exploiter cette source de connaissances pour notamment guider la mise en place d'une stratégie de contrôle. Nous allons voir au cours de cette section qu'un corpus de référence peut être exploité pour à la fois valider les intuitions émises par un expert (modèle de préférences), mais également pour faciliter la détermination des paramètres décisionnels à travers l'observation de phénomènes distinctifs récurrents ou significatifs.

Nous verrons dans un premier temps en quoi un corpus de référence peut nous permettre de compléter la méthodologie de contrôle mise en place. Nous définirons dans un second temps les conditions qu'un corpus doit remplir pour garantir qu'il soit exploitable. Cette section envisage ensuite une perspective intéressante, partiellement étudiée pour le moment, visant à compléter une méthode d'AMCD par des heuristiques statistiques d'interprétation d'un corpus de référence. Bien que ces heuristiques permettent d'évaluer la pertinence du modèle de préférences établi, elles permettent surtout de guider l'expert lors de la formalisation de ses connaissances en tant que modèle de préférences.

4.3.1 Regard statistique sur le modèle de préférences

Les méthodes `ELECTRE III` et `ELECTRE TRI`, en tant qu'approches par surclassement, placent l'expert humain au centre de la définition de la stratégie de décision. Cet aspect constitue le principal avantage de notre démarche et également la raison principale de notre choix pour ce paradigme décisionnel. Ainsi, à travers la formalisation d'un modèle de préférences, l'expert peut modéliser ses connaissances et intuitions sur les informations distinctives à prendre en compte lors du contrôle, mais également sur la façon dont ces données doivent être exploitées. Notamment grâce à l'externalisation de ces informations décisionnelles, le déploiement d'une stratégie de décision devient alors un processus itératif alternant des phases d'expérimentation et de validation. En effet, l'expert peut déterminer un profil décisionnel initial et évaluer le comportement d'une stratégie de contrôle sur un ensemble de phrases tests, puis éventuellement procéder à des modifications ou des spécifications des paramètres décisionnels. L'objectif de cette démarche de construction progressive d'un profil décisionnel est d'atteindre un comportement de contrôle satisfaisant pour un ensemble d'exemples de test.

Le modèle de préférences résultant d'une telle démarche expérimentale provient donc d'une formalisation d'intuitions, c'est-à-dire de connaissances empiriques éventuellement complétées

par une phase inductive de tests sur un ensemble représentatif d'exemples. Il peut cependant apparaître délicat de formaliser ce genre de connaissances empiriques à l'aide de paramètres décisionnels numériques [Mousseau and Slowinski, 1998]. Il est donc nécessaire de se poser la question de la validité d'un tel modèle de préférences empiriques. C'est pourquoi nous avons envisagé de confronter les intuitions d'un expert avec des régularités observées sur un corpus, sous réserve de disponibilité de ce dernier.

Cette démarche nous paraît importante dans la mesure où, comme pour de nombreuses méthodes, qu'elles soient statistiques ou symboliques, la pertinence des valeurs associées à ces paramètres conditionne grandement la qualité des résultats obtenus. En effet, comme nous pourrons le constater dans le chapitre suivant consacré à l'application de nos méthodes sur des cas concrets de contrôle, la pertinence du modèle de préférences utilisé est tout aussi importante que la pertinence des performances de critères associées aux hypothèses concurrentes.

De plus, nous avons également vu que l'un des apports de notre architecture logicielle de contrôle concernait la simplification de l'intégration au cours du processus d'analyse d'informations distinctives sous forme de critères de comparaison. À travers la généralisation de cette notion de critère de comparaison et la centralisation de ces objets décisionnels, il devient plus simple de considérer de nombreux aspects et traits comme des critères de comparaison et de les intégrer dans une procédure de contrôle. Il est cependant à nouveau important d'avoir un regard critique sur l'intérêt et la pertinence des différents critères envisagés lors d'une stratégie de contrôle.

Depuis longtemps, les corpus annotés servent de source de connaissances pour l'extraction de connaissances ou l'évaluation de systèmes. Nous proposons désormais d'envisager un nouveau regard sur l'utilisation de ces corpus afin notamment :

- d'évaluer la pertinence des critères utilisés ;
- de confronter un modèle de préférences expertes avec des observations sur corpus ;
- de suggérer éventuellement des paramètres décisionnels "optimaux" ou plus précisément des paramètres adaptés au corpus utilisé.

Afin d'envisager de tels usages d'un corpus, il convient de s'intéresser à la nature de cette source de connaissances et quelles conditions elle doit remplir pour être exploitée en complément de notre approche de contrôle décisionnel par surclassement.

4.3.2 L'annotation comme l'expression des préférences d'un expert (décideur)

Lors de la présentation des approches par surclassement, nous avons présenté les différentes relations de préférence permettant d'établir une situation de comparaison entre deux hypothèses. Ces relations pouvant spécifier des situations de préférence, d'indifférence ou d'incomparabilité sont désignées par [Mousseau, 2003] comme des préférences orientées "outputs", c'est-à-dire portant sur les hypothèses concurrentes. Le modèle de préférences est au contraire composé de préférences qualifiées d'orientées "inputs", portant non pas sur les hypothèses mais sur les critères qui les qualifient.

Bien qu'il paraisse relativement aisé pour un expert d'établir des relations de préférence entre différentes hypothèses concurrentes déterminant leur pertinence relative, la définition d'un modèle de préférences adapté à un cas de contrôle reste un exercice délicat. À partir de ce constat,

il devient intéressant de déterminer si les préférences orientées "inputs" peuvent être inférées à partir de préférences orientées "outputs". Avant de voir comment ce passage entre les préférences établies sur les hypothèses et les préférences sur les critères peut être envisagé, il apparaît intéressant de décrire dans quel contexte ces préférences orientées "outputs" sont disponibles.

Ces préférences orientées "outputs" correspondent donc à des décisions émises par un expert sur des hypothèses concurrentes. En accord avec ses préférences ou connaissances, ces décisions permettent de statuer sur la pertinence relative des hypothèses concurrentes entre elles. Dans notre contexte de TALN, les corpus de référence peuvent être considérés comme de tels regroupements de décisions. Souvent exploités par des procédures d'apprentissage ou d'évaluation, ces corpus sont constitués d'analyses validées par un expert (ou éventuellement par un croisement des résultats de différents analyseurs). La construction de corpus de référence repose donc sur la validation par un expert des hypothèses valides parmi toutes celles générées par un module de traitement sur un ensemble d'énoncés représentatifs du contexte applicatif. Un tel corpus de référence est donc relatif à un niveau d'analyse (lexical, syntaxique, sémantique) ou une tâche (traduction, identification ou catégorisation des entités nommées, etc.), et comme tout corpus à un domaine d'application (articles de presse, documentations techniques, etc.). Lorsqu'un expert construit un corpus de référence, il exploite ses préférences et ses connaissances pour attester de la validité ou de la non validité des différentes hypothèses concurrentes. Cette validation s'apparente donc à la formulation de préférences orientées "outputs", permettant d'établir notamment une relation de préférence pour les hypothèses valides par rapport à celles jugées comme non valides.

Si désormais l'on considère que le module de traitement à contrôler associe à chaque hypothèse qu'il génère un vecteur de performances décrivant l'évaluation de la pertinence de l'hypothèse sur différents critères, il devient alors possible de reporter les préférences orientées "outputs" (hypothèses) sur les préférences orientées "inputs" (critères). Ainsi, lorsqu'un expert atteste de la validité (resp. de la non validité) d'une hypothèse, il détermine également que le vecteur de performances qui lui est associé est caractéristique d'une hypothèse valide (resp. non valide).

Bien que cette démarche d'annotation ne nous permette pas de reconstituer tous les types de relations (préférence forte, préférence faible, indifférence, incomparabilité) pouvant être établies entre deux hypothèses concurrentes, elle nous apporte tout de même des connaissances intéressantes sur les hypothèses et les critères, qui forment à eux deux les principaux composants d'un problème décisionnel. En effet, cette annotation nous permet de regrouper des hypothèses et leurs vecteurs de performances associés en deux classes : celle des hypothèses valides et celle des hypothèses non valides. Ce passage du marquage des hypothèses concurrentes au marquage de leurs vecteurs de performances respectifs, nous ramène à un contexte classique d'apprentissage supervisé, où nous disposons d'un ensemble d'individus (hypothèses), chacun étant caractérisé par un ensemble d'attributs (vecteur de performances) et d'une classe d'affectation (valide ou non valide). Sur une telle formalisation des connaissances préférentielles apportées par l'annotation d'un expert, il devient alors envisageable d'appliquer des méthodes issues de l'apprentissage automatique supervisé, ou plus généralement de procéder à des observations statistiques. Comme nous allons le voir dans la section 4.3.3, l'application de ces méthodes nous permet d'obtenir des informations intéressantes à la fois sur les hypothèses mais également sur les différents attributs qui les qualifient, telles que la représentativité d'un attribut pour la classe des hypothèses valides, les variations des performances obtenues sur chaque critère pour des hypothèses valides, etc.

Cette classification d'un ensemble d'hypothèses nous permet de dresser ce que nous nommons

une table de performances, qui regroupe les hypothèses concurrentes, leur vecteur de performances et leur classe d'affectation.

Hypohèses	vecteur de performances					annotation
h_1	$g_1(h_1)$	$g_3(h_1)$	...	$g_{m-1}(h_1)$	$g_m(h_1)$	valide
h_2	$g_1(h_2)$	$g_3(h_2)$	...	$g_{m-1}(h_2)$	$g_m(h_2)$	invalide
h_3	$g_1(h_3)$	$g_3(h_3)$	...	$g_{m-1}(h_3)$	$g_m(h_3)$	valide
...	...	...	...	...	...	...
h_{n-1}	$g_1(h_{n-1})$	$g_3(h_{n-1})$	...	$g_{m-1}(h_{n-1})$	$g_m(h_{n-1})$	invalide
h_n	$g_1(h_n)$	$g_3(h_n)$	...	$g_{m-1}(h_n)$	$g_m(h_n)$	invalide

TAB. 4.1 – Tableau de performances construit à partir de l'alignement entre les sorties d'un module de traitement à contrôler et un corpus de référence approprié

Bien que nous ayons développé un logiciel (section 4.3.2) pour faciliter la construction de corpus de référence et donc de tables de performances à partir des hypothèses concurrentes générées par les différents modules qui composent la chaîne de traitement `TiLT`, cette tâche d'annotation est fastidieuse et il n'est pas envisageable de disposer de telles tables de performances pour tous les "points de décision" sur lesquels on souhaite déployer une stratégie de contrôle. Ces tables de performances peuvent cependant être également obtenues en exploitant des corpus de référence "communautaires", comme par exemple le `Penn TreeBank`[11]. En effet, il est possible de procéder à un alignement afin d'annoter automatiquement les hypothèses concurrentes et donc indirectement les vecteurs de performances, en fonction de leur conformité vis-à-vis des hypothèses de référence. Cet alignement est évidemment envisageable uniquement si les hypothèses de référence et les hypothèses concurrentes générées par le module à contrôler sont de même nature, et il faut évidemment que les hypothèses concurrentes aient été générées à partir des énoncés regroupés dans le corpus de référence. Par exemple, le corpus annoté `Penn Tree Bank` peut donc servir de référence pour attester de la validité ou de la non validité des hypothèses générées par le module d'analyse syntaxique de surface (distribution de catégories morpho-syntaxiques) sur des énoncés écrits en anglais.

CorpusTagger

Les différents modules d'analyse qui composent la chaîne de traitement `TiLT` disposent d'une méthode d'affichage permettant de formaliser en `XML` les hypothèses générées. Il est donc possible d'avoir en différents points du processus d'analyse une vue des hypothèses concurrentes manipulées et de l'ensemble des traits et propriétés qui les caractérisent. Par exemple, la méthode d'affichage `XML` du module d'analyse lexicale énumère l'ensemble des unités lexicales construites par accès au lexique ainsi que par l'éventuelle application de méthodes de correction, en spécifiant à la fois les formes fléchies et lemmatisées, les traits lexicaux, phonétiques, morpho-syntaxiques, lexico-sémantiques et d'autres informations de fréquence, etc. Un exemple d'une telle sortie `XML` est proposé en annexe D.1.

Nous avons développé un outil d'annotation, `CorpusTagger`, qui permet de désigner et de marquer les hypothèses correctes et incorrectes parmi l'ensemble des hypothèses concurrentes

[11] `http://www.cis.upenn.edu/~treebank/`

générées. Étant donné la faible lisibilité d'un fichier brut `XML`, l'outil `CorpusTagger` propose d'appliquer des vues (développées sous forme de feuille de transformation `XSLT`) sur ces documents `XML` afin de mettre en exergue certaines caractéristiques et de ne pas afficher des données non-informatives. Nous donnerons un exemple de cette fonctionnalité dans la section 5.1, où l'outil `CorpusTagger` a été utilisé pour constituer un corpus de référence spécialisé sur la tâche en question.

De plus, cet outil spécialisé dans l'annotation de corpus de référence offre des fonctionnalités avancées de marquage afin de simplifier et de rendre plus efficace l'annotation d'une référence. Il est par exemple possible de rechercher l'ensemble des hypothèses contenant un certain patron déterminé par une expression régulière ou une chaîne de caractères surlignée. Il est également possible de reporter l'annotation d'une hypothèse à toutes les hypothèses du corpus contenant des propriétés similaires. Une illustration de cet outil est proposée en annexe C.2.

Cet outil est donc adapté à l'annotation des sorties réalisées par `TiLT` et permet également de générer des tableaux exploitables permettant l'étude des différents critères associés aux hypothèses annotées.

4.3.3 Quelques heuristiques pour suggérer un modèle de préférences à partir d'un corpus de référence

La construction automatique d'un modèle de préférences à partir d'un ensemble d'hypothèses est déjà une pratique connue en AMCD. En effet, sous le terme d'élicitation de préférences, [Mousseau, 2003] propose d'exploiter les techniques de programmation linéaire pour construire un modèle de préférences respectant un ensemble de préférences orientées "inputs" émises par un expert. Ce choix pour la programmation linéaire s'explique par les relations étroites qui relient l'AMCD et la recherche opérationnelle, domaine dans lequel la programmation linéaire excelle. Cependant, comme nous l'avons souligné précédemment en 4.3.2, les tableaux de performances construits à partir des corpus de référence forment un cadre idéal d'application des méthodes issues de l'apprentissage automatique supervisé. Outre le fait d'être motivé par des affinités personnelles avec ce dernier domaine, cette stratégie permet également d'éviter la difficulté soulevée par la non linéarité de la fonction de calcul des indices de surclassement [Mousseau *et al.*, 1999], qui nécessite le recours à une fonction approchée pour la résolution du problème par programmation linéaire.

Quantifier l'importance de chaque critère

À partir d'un tableau de performances, l'une des informations intéressantes qu'il est possible d'extraire concerne l'importance relative des différents critères évalués. En effet, en rappelant que les différentes hypothèses qui composent un tableau de performances peuvent être associées à deux classes différentes : celle des hypothèses valides et celle des hypothèses invalides, on constate qu'identifier l'importance relative de chaque critère revient à évaluer le gain apporté par chacun d'eux lors de l'identification des hypothèses valides. Les méthodes d'apprentissage de métriques répondent particulièrement à ce genre de problématique, en cherchant à évaluer et quantifier la représentativité de différents attributs pour une classe donnée. Parmi les différentes méthodes existantes d'apprentissage de métriques, nous avons utilisé la plus répandue, à savoir la méthode `RELIEF` [Kononenko, 1994]. Bien que cette méthode fut rapidement adaptée aux cas où plusieurs classes sont considérées et où certaines hypothèses sont associées à des vecteurs de per-

formances contenant des valeurs manquantes, nous n'utilisons que la version initiale de RELIEF. Cette méthode est basée sur une mesure de distance permettant d'évaluer la proximité, en terme de performance sur chaque critère, entre les individus de la classe des hypothèses valides et ceux de la classe des hypothèses invalides. Ainsi, plus les performances obtenues par les hypothèses d'une même classe sont homogènes et distantes des performances obtenues par les hypothèses de l'autre classe, plus le critère sera considéré comme représentatif et disposera donc au final d'un poids élevé. Au contraire, si les différences de performances obtenues sur un critère par les hypothèses des deux classes antagonistes ne sont pas importantes, voire contradictoires dans de nombreux cas, ce critère sera considéré comme non discriminant et un poids faible ou négatif lui sera attribué. En effet, la méthode RELIEF estime itérativement un poids pour chaque critère, compris entre -1 et 1 ($\forall w_i \in W, w_i \in [-1, 1]$).

L'algorithme de la méthode RELIEF est illustré ci-dessous, où l'on nomme une table de performances $TP : \{H_1, H_2, ..., H_D\}$ contenant D hypothèses associées à leur vecteur de performances et annotées comme valides ou non valides. On rappelle que $W : \{w_1, w_2, ..., w_m\}$ est le vecteur des poids associés aux m critères utilisés :

Précondition : $TP : \{H_1, H_2, ..., H_D\}$: table de Performances;
Postcondition : $W : \{w_1, w_2, ..., w_m\}$: un vecteur de poids des critères évalués;
$\forall w_i \in W : w_i = 0.0$;
for *i allant de* 1 *à* D **do**
  choisir au hasard un H_i dans TP;
  soit $\underline{H}$ le plus proche voisin de tp_i de la même classe;
  soit $\overline{H}$ le plus proche voisin de tp_i de l'autre classe;
  for *j allant de* 1 *à m (i.e. pour chaque critère)* **do**
    $w_j := w_j - \text{diff}(g_j(H_i), g_j(\underline{H}))/D + \text{diff}(g_j(H_i), g_j(\overline{H}))/D$;
  end
end

Algorithme 3 : Algorithme RELIEF : estimation des poids des critères

L'algorithme RELIEF illustré précédemment (Algo. 3) est donc basé sur une mesure de distance pour identifier les plus proches voisins des hypothèses sélectionnées itérativement de manière arbitraire. Nous avons utilisé une distance de Manhattan pour mesurer la distance entre deux vecteurs de performances.

$$Distance(h_i, h_j) = \sum_{k=1}^{m} |(g_k(h_j) - g_k(h_i))|$$

L'application de cette simple métrique de distance nécessite évidemment que les différentes performances de critères soient commensurables. Bien que ceci ne soit pas initialement le cas, nous procédons à une normalisation de toutes les valeurs du tableau de performances avant l'application de l'algorithme RELIEF. Cette normalisation sur l'espace $[0, 1]$ est tout simplement obtenue en rapportant chaque performance par rapport à la performance maximale contenue dans le tableau de performances sur le critère concerné.
Les poids estimés par la méthode RELIEF sont ainsi normalisés sur $[-1, 1]$, ce qui nous permet d'identifier facilement les critères jugés inutiles ou plus précisément non discriminants vis-à-vis de la classe des hypothèses valides, c'est-à-dire ceux dont le poids final est négatif. Comme nous le verrons par la suite, outre le fait de disposer d'une distribution de poids particulièrement représentative d'un corpus de référence, l'interprétation des résultats de l'algorithme RELIEF nous

permet d'identifier les critères n'apportant pas d'information très discriminante et ainsi de valider ou d'invalider les intuitions de l'expert. Nous pouvons en effet comparer le choix et la hiérarchie des critères établis par l'expert avec les résultats de l'algorithme `RELIEF`.

Cette démarche n'est pas une extension originale des méthodes d'AMCD. Il apparaît en effet évident qu'il est plus simple pour un expert de prendre des décisions plutôt que d'expliquer le processus lui ayant permis d'atteindre ces décisions. De nombreux travaux ont été menés afin d'établir un lien entre les décisions d'un expert, que Mousseau [Mousseau, 2003] qualifie de préférences orientées "outputs", c'est-à-dire portant sur les hypothèses, et les paramètres décisionnels, qualifiés de préférences orientées "inputs", permettant à une méthode d'AMCD d'arriver à ces décisions. Ainsi, [Mousseau *et al.*, 2001] et [Benabbou *et al.*, 2004] ont proposé d'exploiter des préférences orientées "outputs" établies par un expert pour inférer les poids des critères utilisés pour respectivement les méthodes `ELECTRE` en général, `ELECTRE TRI` et `PROAFNT` (voir [Mousseau, 1995] et [Roy and Mousseau, 1996] pour une étude plus théorique et méthodologique de cette démarche). Ces deux méthodes exploitent la programmation linéaire pour déterminer une distribution de poids à affecter aux différents critères utilisés dans le problème décisionnel. Bien que les poids calculés soient rigoureusement fidèles aux préférences orientées "outputs" proposées en entrée des algorithmes, ces approches nécessitent que les autres paramètres décisionnels (seuils et coupes) soient définis *a priori*. Le choix des valeurs à affecter à ces autres paramètres décisionnels affecte grandement la distribution de poids obtenue lors de cette phase d'élicitation.

En revanche, l'utilisation d'une méthode d'apprentissage de métrique nous permet de générer une distribution de poids représentant au mieux le corpus de référence (i.e. d'apprentissage), et n'exploitant que le tableau de performances comme données d'entrée. De plus, l'algorithme déployé par la méthode `RELIEF` est indépendant de la méthode d'AMCD qui exploitera la distribution de poids estimée. Ainsi, la distribution de poids issue de l'application de l'algorithme `RELIEF` peut être utilisée en tant que vecteur de poids des critères, qu'il s'agisse d'une problématique de classement ou de sélection (méthode `ELECTRE III`) ou d'une problématique de tri (méthode `ELECTRE TRI`).

Identifier des zones de préférences pour suggérer les paramètres décisionnels

Une fois les critères pertinents identifiés par l'application de la méthode `RELIEF` sur le tableau de performances et leur importance relative mesurée, nous aurions pu exploiter la programmation linéaire pour suggérer les autres paramètres décisionnels (seuils et coupe) en utilisant notamment les travaux de [Dias and Mousseau, 2006], [Ngo The and Mousseau, 2002] et [Mousseau, 2003], où l'identification des autres paramètres est basée sur une formalisation en tant que problème d'optimisation linéaire. Par manque de temps notamment, mais également afin d'être cohérent avec la méthode présentée précédemment, nous avons mis en place des heuristiques basées sur une étude statistique des tableaux de performances visant à suggérer des valeurs possibles aux autres paramètres décisionnels. Ces heuristiques ont pour unique objectif de fournir une aide aux experts lors de la mise en place d'une stratégie de contrôle et donc d'un modèle de préférences. Cette aide est basée sur l'identification de valeurs de performances pouvant représenter des situations particulières de préférences, d'indifférence ou de veto entre les hypothèses comparées. En aucun cas, ces heuristiques ne peuvent être considérées comme des méthodes fournissant des valeurs optimales pour des paramètres décisionnels, mais bien comme des suggestions de valeurs possibles ou alors des indices sur des zones de valeurs pouvant être particulièrement intéressantes dans le domaine de définition des critères évalués.

Pour identifier les zones potentielles de préférence et d'indifférence, nous proposons d'étudier et de comparer la répartition des hypothèses validées et invalidées résultant de l'alignement entre les sorties d'un module de traitement et un corpus de référence approprié sur le domaine de définition de chacun des critères concernés. À partir de cette répartition illustrée par la figure 4.9, nous exploitons des heuristiques matérialisées par des seuils pour déterminer des intervalles pouvant correspondre à des zones de préférence, d'indifférence et de veto. Pour chaque valeur du domaine de définition du critère, nous calculons la proportion d'hypothèses valides et invalides supérieures à cette valeur.

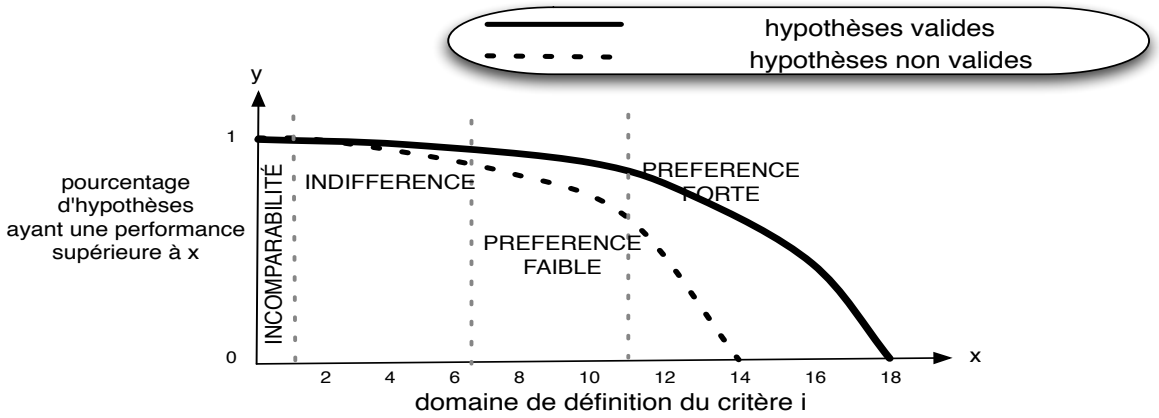

FIG. 4.9 – Répartition des hypothèses et heuristiques d'identification des zones de préférence, d'indifférence et de veto.

Ainsi, pour chaque critère contenu dans le tableau de performances, de telles courbes de répartition sont établies. Nous proposons ensuite d'utiliser différents seuils établis empiriquement pour déterminer des intervalles de valeurs dans le domaine de définition du critère pouvant correspondre à des zones de préférence, d'indifférence et de veto :

- zone d'indifférence :
  Rappelons que l'indifférence témoigne d'une situation où il est délicat de statuer de la domination d'une hypothèse par rapport à une autre. Nous proposons de considérer l'intervalle pour lequel la valeur absolue de la différence entre les proportions d'hypothèses valides et invalides est inférieure à 5%.
- zone de préférence forte :
  Symétriquement à la situation précédente, la zone de préférence forte correspond à l'intervalle où la proportion d'hypothèses valides est significativement supérieure (plus de 15%[10]) à la proportion d'hypothèses invalides.
- zone de préférence faible :
  Correspond à une zone intermédiaire entre l'indifférence et la préférence forte.
- zone d'incomparabilité :
  Cette zone est délimitée par une valeur du domaine de définition du critère concerné en dessous de laquelle on ne trouve que des hypothèses non valides.

[10]Ces seuils ont été définis de manière empirique et nécessiteraient une étude de sensibilité pour être plus pertinents.

Il est ensuite possible d'exploiter ces intervalles pour suggérer des valeurs possibles aux différents paramètres décisionnels dont la sémantique est justement d'établir de telles situations de préférence, d'indifférence et de veto. Cependant, ces paramètres ou plus précisément ces seuils ne sont pas utilisés de la même façon par les méthodes de contrôle présentées précédemment (section 3.2.3) en fonction notamment de la problématique visée. En effet, pour la problématique de tri, c'est-à-dire la méthode `ELECTRE TRI`, les seuils de préférence, d'indifférence et de veto permettent d'appréhender la plus ou moins grande supériorité d'une performance obtenue par une hypothèse sur un critère par rapport à une limite d'acceptabilité d'une classe sur ce même critère. Alors que pour les problématiques de sélection ou de classement (`ELECTRE III`), ces seuils déterminent le rapport de supériorité entre deux hypothèses sur un même critère de comparaison. Dans le premier cas, il s'agit donc de comparer une valeur pouvant varier (la performance de l'hypothèse) avec une valeur fixe (la limite d'acceptabilité d'un critère pour une classe). Dans le second cas, les seuils mesurent des écarts entre des variables (les performances des hypothèses comparées). À partir de ces constats, nous allons constater que l'interprétation de la répartition des hypothèses valides et non valides dans la table des performances n'est pas identique pour ces deux problématiques.

Si l'on considère une répartition des hypothèses valides et non valides pour un critère i telle que celle illustrée par la figure 4.9, les seuils suggérés dans le cadre d'une problématique de tri `ELECTRE TRI` sont les suivants :

- limite d'acceptabilité l_i :
  Cette limite a été établie au centre de la zone d'indifférence, c'est-à-dire le milieu de l'intervalle délimité par la performance minimale obtenue par une hypothèse valide dans le tableau de performances et le début de la zone de préférence faible.
- seuil d'indifférence q_i :
  Correspond à la borne inférieure de la zone de préférence faible, c'est-à-dire "l'intersection" des deux courbes. Les performances des hypothèses étant comparées avec la limite d'acceptabilité, nous établissons le seuil d'indifférence $q_i := q\prime - l_i$.
- seuil de préférence p_i :
  Correspond à la borne inférieure de la zone de préférence forte, c'est-à-dire la première valeur du domaine de définition du critère i où la proportion d'hypothèses valides est nettement supérieure (plus de 10%[11]) à la proportion d'hypothèses invalides. Les performances des hypothèses étant comparées avec la limite d'acceptabilité, nous établissons le seuil d'indifférence $p_i := p\prime - l_i$.
- seuil veto v_i :
  Correspond à la différence entre la limite d'acceptabilité et la performance minimale associée à une hypothèse valide dans la table de performances sur le critère concerné : $v_i = l_i - v\prime$.
  Afin de respecter les contraintes liées à la validité d'un modèle de préférence, nous vérifions notamment que $p_i \geq v_i$. Si cette condition n'est pas initialement respectée, nous prenons la première valeur $v\prime\prime$ comprise entre 0 et $v\prime$ qui respecte cette condition ($l_i - v\prime\prime \geq v_i$).

La figure 4.10 illustre donc la disposition de ces seuils suggérés par rapport à une répartition des hypothèses valides et non valides sur un domaine de définition d'un critère.

Il est évident que la répartition des hypothèses sur le domaine de définition du critère étudié ne suit pas forcément des variations si nettes, permettant d'identifier distinctement des zones de préférence, d'indifférence et de veto. La figure 4.11 illustre par exemple un cas où les situations de

[11] Ces seuils ont été définis de manière empirique et nécessiteraient une étude de sensibilité pour être plus pertinents.

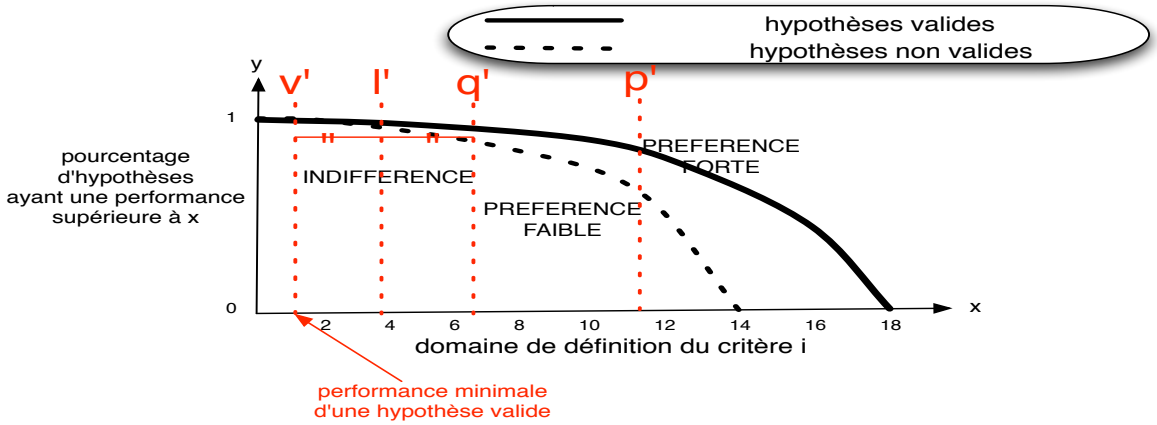

FIG. 4.10 – Identification des seuils suggérés par rapport aux zones de préférence, d'indifférence et de veto

comparaison ne peuvent être facilement différenciées. Ce cas particulier revient à considérer que $q_i = p_i$, ce qui ne constitue pas un problème pour l'application d'une méthode par surclassement.

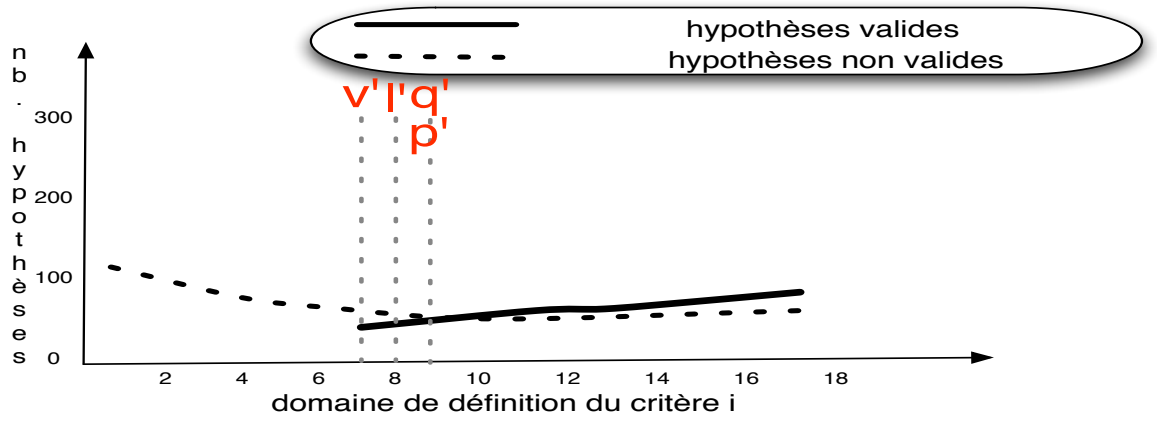

FIG. 4.11 – Un exemple de répartition particulière des hypothèses valides et non valides

Pour les problématiques de classement et de sélection (`ELECTRE III`), les seuils ne correspondent pas à des limites telles que nous les avons directement déterminées à partir des courbes de répartition, mais à des différences. Nous modifions donc dans un premier temps les courbes de répartition des hypothèses, afin que le rapport entre les hypothèses valides et non valides puisse être établi non pas en terme de performance "brute" sur le domaine de définition d'un critère mais en terme d'écart. Ainsi, pour chaque écart possible dans le domaine de définition du critère étudié, nous établissons le rapport entre le nombre d'hypothèses valides et le nombre d'hypothèses non valides dont la performance est au moins supérieure à cet écart. Pour que ce rapport soit pertinent, il faut évidemment rapporter ce nombre au nombre total d'hypothèses valides et non valides contenues dans la table de performances.

Par exemple, si l'on considère un critère i, un nombre V d'hypothèses valides et un nombre N d'hypothèses non valides dans le tableau de performances, le rapport entre les hypothèses valides et non valides pour un écart de 2 minimum sera calculé de la façon suivante :

$$rapport(i,2) = \frac{\frac{|HV_{(i,2)}|}{V}}{\frac{|HN_{(i,2)}|}{N}}$$

où, $HV_{(i,j)}$ est l'ensemble des hypothèses valides $h_k \in TP$ telles qu'il existe une hypothèse non valide $\exists h_l \in TP$ telle que $g_i(h_k) - g_i(h_l) \geq j$
et où $HN_{(i,j)}$ est l'ensemble des hypothèses non valides $h_l \in TP$ telles qu'il existe une hypothèse valide $\exists h_k \in TP$ telle que $g_i(h_l) - g_i(h_k) \geq j$

À l'aide de la méthode de calcul précédente, nous pouvons établir une courbe retraçant le rapport des écarts entre HV et HN, comme l'illustre la figure 4.12.

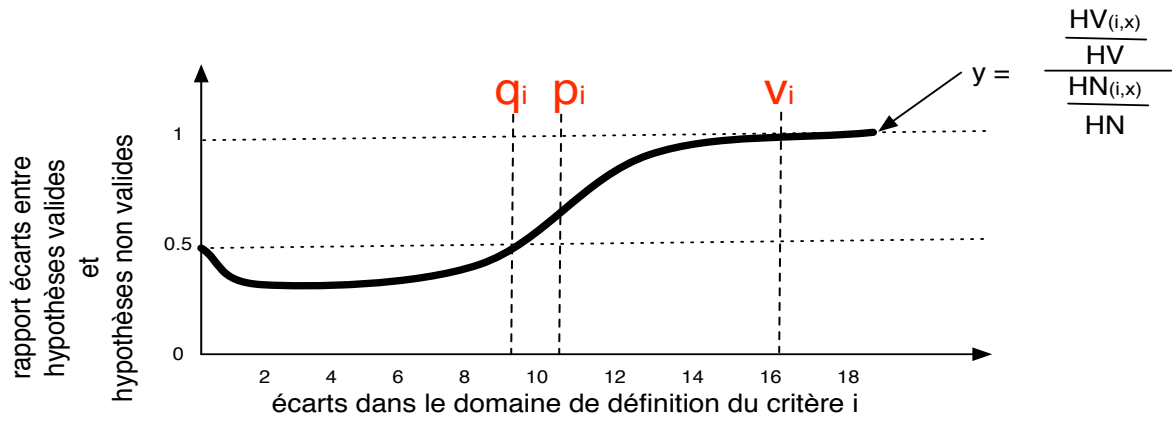

FIG. 4.12 – Répartition des écarts entre les hypothèses valides et non valides et identification des seuils

À partir de cette courbe de répartition des écarts, nous identifions des valeurs possibles pour les différents seuils en utilisant les heuristiques suivantes et comme l'illustre la figure 4.12 :

- seuil d'indifférence q_i :
  Correspond au premier écart où le rapport entre les hypothèses valides et non valides est supérieur à $0,5$.
- seuil de préférence p_i :
  Correspond au premier écart où le rapport entre les hypothèses valides et non valides est supérieur à $0,6$[12].
- seuil veto v_i :
  Correspond au premier écart où le rapport entre les hypothèses valides et non valides est égal à 1.

Lors de la présentation des méthodes `ELECTRE III` et `ELECTRE TRI` nous avons introduit un paramètre technique supplémentaire, le seuil de coupe, qui détermine la limite de validation d'une relation de surclassement à partir de l'indice de crédibilité de surclassement. Nous avons

[12] De même que pour les heuristiques précédentes, ce seuil de 10% défini empiriquement nécessiterait une étude de sensibilité pour être plus pertinent

également cherché à exploiter les tables de performances pour suggérer une valeur à ce paramètre qui soit représentative des hypothèses de référence disponibles. Nous avons vu que les heuristiques décrites lors de la section 4.3.3 nous permettaient de suggérer à l'expert un modèle de préférences complet (poids et seuils de préférence, d'indifférence, de veto) et également de déterminer des valeurs possibles pour l'élaboration d'un profil de classe permettant de discriminer les hypothèses positives des hypothèses négatives. À partir de ces éléments, nous calculons pour entre chaque hypothèse des tables de performances et le profil de la classe des hypothèses valides un indice de crédibilité de surclassement. Cet indice de crédibilité variant sur l'intervalle $[0, 1]$, nous déterminons ensuite la valeur, toujours comprise dans cet intervalle, qui sépare au mieux les hypothèses positives des hypothèses négatives. Cette valeur est ensuite suggérée comme seuil de coupe. Cette recherche par essais successifs s'effectue entre l'indice de crédibilité minimal obtenu par un exemple positif et l'indice de crédibilité maximal obtenu par un exemple négatif, comme l'illustre la figure 4.13.

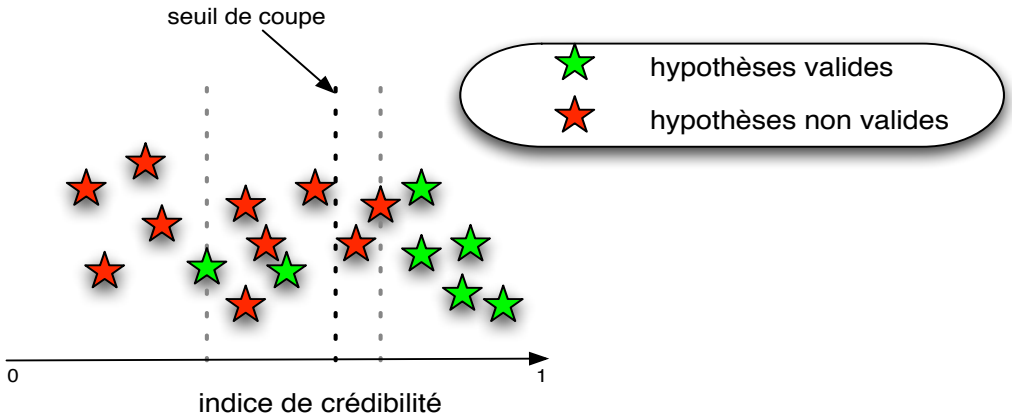

FIG. 4.13 – Identification du seuil optimal de coupe pour la séparation des exemples d'apprentissage positifs et négatifs

Une ouverture vers des perspectives de contrôle

On constate aisément que les différentes heuristiques proposées pour suggérer des valeurs possibles aux seuils de préférence, d'indifférence et de veto associés aux critères ne semblent efficaces que pour des situations particulières de répartition des hypothèses sur les domaines de définition des critères. Il est cependant important de souligner à nouveau que ces derniers travaux ne visent pas à produire un modèle de préférences "optimal" à partir d'un corpus de référence. Ces heuristiques ont pour objectif principal d'aider et éventuellement de guider un expert (décideur) lors de l'élaboration d'un modèle de préférences en lui suggérant par exemple des zones potentiellement charnières pour la comparaison des hypothèses.

Ces différentes heuristiques permettent tout de même d'améliorer la compréhension de l'expert sur l'intérêt des différents critères utilisés. En effet, le cas particulier où aucune des zones de préférence, d'indifférence et de veto n'a pu être établie constitue en soi une information intéressante car il permet à l'expert d'identifier des critères potentiellement mal-définis ou inutiles.

Lors des chapitres précédents, nous avons souligné l'originalité de notre approche de contrôle en insistant sur l'intersection créée entre le TALN et l'AMCD. Avec ces dernières perspectives visant à exploiter les corpus de référence pour l'aide à la construction de modèles de préférences, nous introduisons un troisième domaine dans nos recherches, celui de l'étude statistique de corpus notamment à l'aide de méthodes issues de l'apprentissage supervisé. Bien que cette dernière perspective nécessite et mérite d'être approfondie davantage, nous avons tout de même mis en avant une nouvelle méthode d'exploitation des corpus de référence.

Ces derniers travaux sur l'étude statistique des performances de critères complètent les fonctionnalités proposées par le module de contrôle décisionnel qui a été intégré dans la chaîne de traitement linguistique `TiLT`. En effet, nous avons vu que ce module simplifiait énormément l'intégration et la définition de critères de comparaison sur les hypothèses générées par les modules de traitement. Nous proposons désormais d'évaluer partiellement la pertinence et la validité de ces informations et ainsi de proposer un retour aux intuitions émises par l'expert lors de la mise en place d'une stratégie de contrôle.

Conclusion de chapitre

Notre démarche de contrôle basée sur une approche par surclassement s'inscrit l' optique de concevoir un système à la disposition des experts en charge du déploiement d'instances de `TiLT` pour les différents contextes d'analyse. Nous répondons ainsi à l'observation formulée par [Bachimont, 1992] (page 62) selon laquelle un module de contrôle doit fournir une médiation et des interactions entre une architecture et les connaissances de contrôle prises en compte pour résoudre un problème.

5

Évaluation de l'apport de l'AMCD pour le contrôle d'une chaîne de TALN

Sommaire

5.1 Classement des rubriques métier pour une requête soumise à un service d'annuaire de professionnels 103
5.1.1 Le processus d'indexation de TiLT et l'apparition des indéterminations . 103
5.1.2 Un "point de décision" déjà établi 104
5.1.3 Vers une meilleure exploitation des critères de comparaison 106
5.1.4 Résultats et interprétations 108
5.2 Identification des cas de reprise anaphorique : problématique de tri 110
5.2.1 Présentation du problème 111
5.2.2 Présentation de la démarche expérimentale 116
5.2.3 Interprétation des résultats : propriétés linguistiques des reprises anaphoriques 127
5.3 Quelle transcription pour un SMS ? problématique de sélection . . 130
5.3.1 Le processus symbolique de transcription des SMS 130
5.3.2 Caractérisation et quantification des indéterminations 133
5.3.3 Vers un contrôle statistique du processus d'analyse 139
5.3.4 Application de la stratégie de contrôle, résultats et interprétations . . . 145

Lors du chapitre précédent, nous avons vu comment notre proposition de contrôle du processus d'analyse linguistique par des méthodes issues de l'AMCD avait été implantée dans la chaîne de traitement `TiLT`. La description détaillée de cette démarche de conception et d'implémentation d'un système complet et générique de contrôle nous a permis de valider la faisabilité de cette proposition. Au cours de ce chapitre, nous allons nous focaliser sur l'apport de notre système à travers le contrôle de trois cas précis d'indétermination observés lors de l'application de `TiLT` sur des contextes d'analyse différents.

À travers cette étape indispensable d'évaluation, nous cherchons principalement à valider trois aspects :

1. l'utilisabilité du module de contrôle :
   Est-ce que les propriétés de généricité et de simplification que nous avons mises en avant lors de la conception de la méthodologie de contrôle sont vérifiées ? Ce qui revient à évaluer l'utilisabilité du module de contrôle, à la fois pour le développeur qui souhaite intégrer une étape de contrôle dans le code d'un module de traitement, mais également pour un expert, par exemple un linguiste, qui cherche à mettre en place un modèle de préférences adapté à un contexte de contrôle précis.
2. la pertinence et la traçabilité des décisions émises :
   Évaluer la pertinence des décisions émises à la suite d'une étape de contrôle revient à vérifier si on résout, au moins partiellement, notre problématique liée à la propagation d'hypothèses erronées. Comme nous l'avons signalé dans la section 2.1.2, la diminution de la proportion d'hypothèses erronées propagées se mesure à l'aide d'une métrique de précision. Nous pourrons ainsi comparer la précision obtenue entre l'application d'un processus d'analyse non contrôlé et celle obtenue suite à la mise en place d'une stratégie de contrôle décisionnelle. Nous verrons également en quoi la centralisation des informations décisionnelles nous offre une meilleure traçabilité des processus de contrôle et de leurs résultats.
3. l'apport d'une approche par surclassement pour notre problématique de contrôle :
   Lors de la présentation de nos objectifs et de notre hypothèse de recherche (section 2.3.3), nous avons expliqué que notre choix s'était porté sur les méthodes d'AMCD par surclassement pour principalement quatre raisons :
   (a) le fait que ces méthodes soient spécialisées dans la résolution de problèmes décisionnels ;
   (b) la prise en compte de connaissances expertes ;
   (c) l'indépendance vis-à-vis de corpus d'apprentissage, c'est-à-dire d'être utilisable en ne reposant que sur des connaissances expertes.
   (d) la traçabilité et l'interprétabilité du processus de décision.

   Il sera donc finalement intéressant de voir si ces différentes motivations apparaissent réellement comme des avantages validant l'utilisation de ce paradigme décisionnel pour notre problématique.

La chaîne de traitement `TiLT` est composée d'un ensemble de modules d'analyse (section 1.2.2) pouvant être paramétrés et appliqués séquentiellement afin de constituer différents processus de traitement applicables dans de nombreux contextes applicatifs. Chacun des modules intégrés dans un processus de traitement constitue potentiellement un cas d'indétermination et donc d'expérimentation pour notre module de contrôle. Ainsi, parmi les nombreux cas d'expérimentation envisageables, nous avons décidé de nous intéresser à trois cas particuliers :

1. l'indexation de requêtes soumises à un annuaire en ligne de professionnels ;
2. l'identification des reprises anaphoriques ;
3. et la transcription de SMS.

Bien que nous verrons lors des perspectives que l'utilisation de notre module de contrôle a été envisagée pour d'autres cas d'indétermination, ce chapitre est focalisé sur la présentation de ces trois expérimentations. Le choix pour ces trois cas d'expérimentation s'explique tout d'abord par notre volonté de travailler sur les mêmes domaines applicatifs que le reste de l'équipe, mais également par le fait que nous disposons de corpus d'évaluation et/ou d'apprentissage spécifiques à ces tâches. De plus, nous constaterons que ces trois cas d'expérimentation nous ont permis d'appliquer les trois différentes problématiques soulevées par la mise en place d'une stratégie de contrôle, à savoir le classement, la sélection et le tri. Les trois sections qui composent ce chapitre décrivent donc ces contextes d'indétermination, la procédure d'application de notre méthode de contrôle et les résultats obtenus.

5.1 Classement des rubriques métier pour une requête soumise à un service d'annuaire de professionnels

La chaîne de traitement `TiLT` est exploitée depuis plusieurs années pour une indexation de requêtes soumises à un service d'annuaire en ligne de professionnels. À travers une description de ce processus d'analyse, nous allons constater que le problème de la génération d'hypothèses d'indexation concurrentes a déjà été soulevé et que ce "point d'embarras" a déjà été transformé en "point de décision". L'objectif de cette première expérimentation consiste alors à déterminer si une approche par surclassement permet une exploitation plus fine des critères disponibles pour comparer les hypothèses d'indexation concurrentes.

5.1.1 Le processus d'indexation de `TiLT` et l'apparition des indéterminations

Une instance particulière de la chaîne de traitement `TiLT` a été mise au point pour assister le fonctionnement d'un annuaire en ligne de professionnels, et plus précisément pour l'interprétation des requêtes soumises par les utilisateurs. `TiLT` est donc utilisé par ce service pour effectuer une indexation des requêtes visant à identifier une catégorie de professionnels que nous nommons `Rubrique`.

L'utilisation de technologies de TALN dans ce contexte applicatif permet d'obtenir une meilleure robustesse du processus d'indexation et propose ainsi une plus grande liberté aux utilisateurs dans la formulation de leurs requêtes. À travers une analyse de la construction des requêtes, `TiLT` permet notamment de détecter les mots ou expressions se référant à une catégorie de professionnels ("hotel" dans "hotel ibis avec air climatisé et piscine orchidée"), ceux apportant un complément ou une précision sur les professionnels recherchés ("air climatisé" et "piscine" pour cette même requête) et ceux correspondant à du bruit (orchidée).

Dans ce contexte d'indexation, `TiLT` propose également une fonctionnalité intéressante de correction automatique des requêtes mal-orthographiées.

Afin de remplir toutes ces fonctionnalités, le processus d'analyse s'appuie sur l'activation de quatre modules de traitement appelés séquentiellement. Comme l'illustre la figure 5.1, les requêtes sont tout d'abord segmentées et chaque segment identifié est soumis à une analyse lexicale. Si la forme analysée n'est pas présente dans le lexique, différents modes de correction sont activés

(troncature, correction morphologique, correction typographique). Cette phase d'analyse lexicale est complétée par une identification des locutions (mots composés tels que "machine à laver").

Les unités lexicales construites lors de l'application du module d'analyse lexicale sont ensuite soumises à un traitement syntaxique de surface. Les connaissances syntaxiques apportées par ce traitement permettent de catégoriser les groupes de mots (noms communs, verbes, etc.) et ainsi d'identifier les mots désignant des rubriques et ceux désignant des compléments sur les rubriques.

Le processus d'analyse s'achève par une identification d'expressions désignant des entités telles que les noms propres ("Pierrette Bouchaud"), les enseignes ("Le livre en folie") et certains compléments ("vue sur mer").

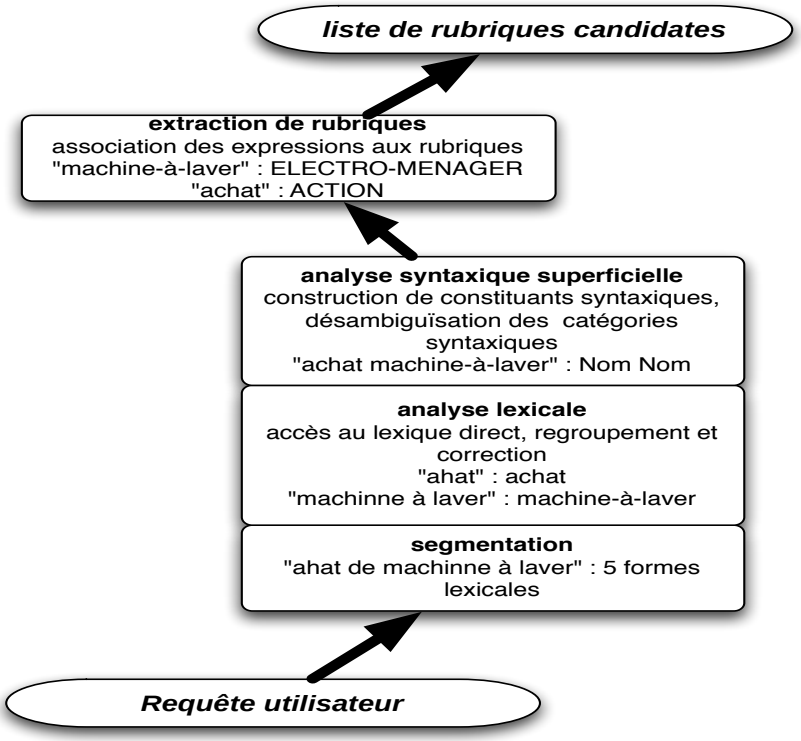

FIG. 5.1 – Processus d'analyse linguistique réalisé par `TiLT` pour l'indexation des requêtes

5.1.2 Un "point de décision" déjà établi

Le processus d'analyse présenté précédemment permet d'identifier les mots ou groupes de mots qui doivent être ensuite utilisés pour activer des rubriques métier et ceux devant être utilisés comme compléments pour la sélection des professionnels. Cette dernière partie n'est pas prise en charge par `TiLT` mais par la société hébergeant le service d'annuaire.

Une fois les mots pertinents pour l'indexation identifiés, un processus d'extraction des rubriques est lancé. Il arrive cependant, dans le cas de requêtes complexes (basées sur plusieurs mots), que plusieurs rubriques soient activées. La figure 5.2 illustre les différentes rubriques activées par la requête utilisateur pourtant simple "salon de coiffure".

Ainsi, le processus d'indexation que nous venons de présenter se trouve dans la plupart des cas confronté à la gestion d'un ensemble d'hypothèses concurrentes. Cet état correspond à ce que

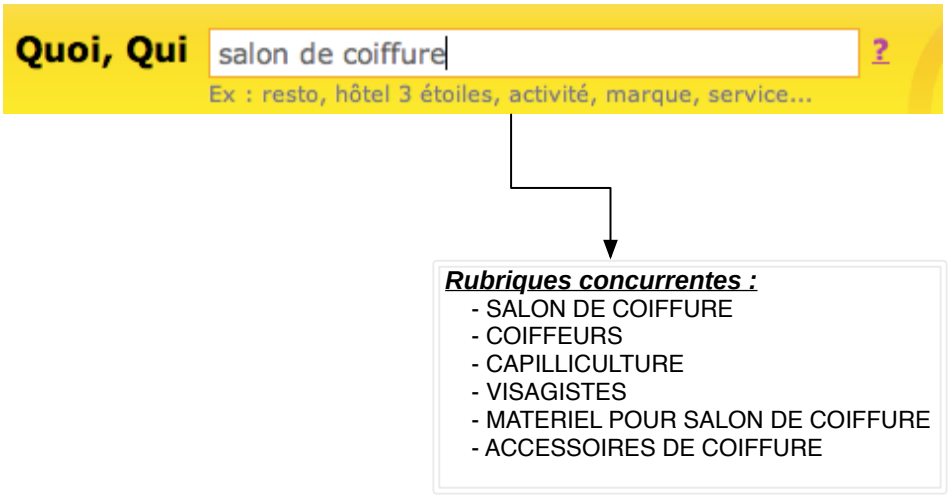

FIG. 5.2 – Rubriques métiers associés à la requête "salon de coiffure"

nous avons désigné sous le terme de "point d'embarras".

Face à la nécessité de disposer d'une évaluation de la pertinence relative des rubriques concurrentes, différentes connaissances supplémentaires ont été intégrées pour traiter cette association entre une requête et de multiples rubriques concurrentes. Ces connaissances supplémentaires visent à donner sous différents aspects une évaluation de la pertinence de chacune des rubriques proposées. La mise à disposition de ces informations permet alors de transformer le "point d'embarras" observé en un "point de décision".

Une analyse des résultats générés par `TiLT` en tant que système d'indexation et des attentes du service d'annuaire de professionnels a permis aux experts linguistes d'identifier quatre connaissances supplémentaires pouvant être utilisées pour l'évaluation de la pertinence des rubriques générées :

1. un score ;
   Correspond au nombre de mots de la requête ayant contribué à activer une rubrique.
2. un coefficient de redondance ;
   Ce coefficient complète et minimise le score et correspond au nombre de mots proches sémantiquement et activant une même rubrique sans pour autant apporter de complément (par exemple, chaque mot de la requête "voiture automobile véhicule" active une même rubrique "VEHICULE" ayant ainsi un score et également une redondance élevés.
3. un rang ;
   Représente la similarité sémantique (exploitation d'un thésaurus) entre les mots de la requête et la rubrique activée.
4. un coefficient d'homogénéité.
   Permet d'exprimer la cohérence entre les mots de la requête activant une rubrique et ceux apportant un complément pertinent pour la rubrique.

Ces quatre informations distinctives ajoutées au processus initial d'indexation vise donc à transformer le "point d'embarras" formé par la génération de rubriques concurrentes en un "point

de décision". Cette situation répond donc aux spécifications d'un problème décisionnel tel que nous l'avons présenté dans la section 3.1.2, où les informations distinctives décrites précédemment forment les critères de comparaison exploités pour traiter ce problème décisionnel.

Le score, le rang et le coefficient d'homogénéité correspondent par nature à des fonctions strictement croissantes et peuvent ainsi être directement considérés comme des critères de comparaison exploitables par notre approche de surclassement (voir sections 2.2.4 et 3.1.2). Ce qui signifie que plus la performance obtenu par une hypothèse sur ces critères est élevée plus l'hypothèse sera jugée comme pertinente. En revanche, le coefficient de redondance, agissant comme une pénalité par rapport au score, doit être minimisé lors de la construction du classement. Une inversion de son sens de variation a donc été nécessaire afin de le considérer comme un critère.

Pour chacune des rubriques générées par le processus d'indexation de `TiLT`, une performance est calculée sur ces différents critères d'évaluation. Ces performances sont ensuite exploitées pour évaluer la pertinence relative des différentes hypothèses concurrentes et établir un classement complet des rubriques.

Dans la version initiale du processus d'indexation, le classement entre les rubriques concurrentes est effectué à l'aide d'une méthode lexicographique. Les experts ont ainsi établi des relations d'importance entre ces quatre critères. Les différentes rubriques sont dans un premier temps classées selon leur score. En cas d'*ex-aequo*, le coefficient de redondance est exploité, puis si des hypothèses doivent encore être départagées le rang et finalement le coefficient d'homogénéité sont utilisés. Comme nous l'avons souligné dans la section 3.1.3 consacrée à la présentation des différentes méthodes multicritère ces méthodes entraînent fréquemment l'exploitation d'un seul critère, perdant ainsi les principaux atouts d'une approche multicritère tels que la complémentarité ou la gestion des contradictions. En effet, la stratégie actuelle de classement privilégie une rubrique possédant un score très légèrement supérieur à celui des autres hypothèses même si ce score reflète une forte redondance et une faible homogénéité.

5.1.3 Vers une meilleure exploitation des critères de comparaison

Le processus d'indexation réalisé par `TiLT` conduit donc à un cas typique d'indétermination caractérisé par l'association de plusieurs rubriques candidates pour une requête initiale. La prise en compte de connaissances supplémentaires matérialisées par quatre critères permet d'obtenir un cadre de contrôle intéressant, dont l'objectif, fixé par le service d'annuaire, est d'établir un classement des différentes rubriques candidates.

Au cours de cette première expérimentation, nous avons cherché à évaluer l'apport de notre approche de contrôle sur la problématique de classement des rubriques concurrentes d'indexation. Les objectifs fixés dans ce cas précis de contrôle sont doubles. Dans un premier temps, il s'agissait de confirmer la simplicité et l'efficacité d'usage de notre module de contrôle à travers la déclaration et la centralisation des critères associés aux rubriques générées, puis à leur exploitation au cours de l'application de l'opérateur de contrôle par surclassement dédié aux problématiques de classement. Dans un second temps, une comparaison des résultats obtenus par les deux méthodes de classement nous permettra de déterminer si la prise en compte systématique de l'ensemble des critères de comparaison disponibles conduit à de meilleurs résultats.

Afin de mener cette évaluation, nous avons tout d'abord procédé à l'élaboration d'un corpus de référence, à partir de deux corpus "bruts" de requêtes. Le premier corpus regroupe les cent requêtes les plus fréquemment soumises à l'annuaire en ligne de professionnels. Le second correspond à un ensemble de 1000 requêtes à la fois moins fréquentes et plus complexes (voir l'annexe F.1 pour des exemples de requêtes).

À l'aide de l'outil `CorpusTagger` (voir annexe C.2) que nous avons développé, le linguiste en charge de ce projet, en tant qu'expert du domaine, a été sollicité pour annoter les différentes rubriques concurrentes proposées par `TiLT`. L'annexe F.2 propose un exemple de la sortie `XML` générée par `TiLT` et correspondant également aux données visualisées et annotées à l'aide de `CorpusTagger`. Nous avons constaté que ce travail d'annotation était très délicat dans la mesure où il était difficile, en tant que personne non spécialisée dans ce cas particulier d'indexation, de se substituer à un expert du domaine[13]. En effet, pour la plupart des requêtes, l'expert attend des réponses bien précises et l'évaluation de la pertinence des différentes rubriques est difficile à établir à partir des distinctions souvent faibles pouvant exister entre les hypothèses concurrentes. Ainsi, malgré la simplification apportée par `CorpusTagger` pour cette tâche d'annotation notamment à l'aide d'une visualisation plus naturelle et agréable des informations présentes dans les fichiers `XML` et le report d'annotation sur des rubriques possédant de nombreuses similarités, seulement 157 requêtes ont au final été annotées (les 100 plus fréquentes et un échantillon de 57 requêtes diverses). Sur ces 157 requêtes, 1208 rubriques ont été désignées comme correctes, acceptables ou incorrectes. On constate que sur ces 157 requêtes, le processus d'indexation réalisé par `TiLT` génère en moyenne 12 rubriques concurrentes par requête.

Les mesures fréquemment utilisées en `TALN` pour évaluer l'efficacité des processus, la précision, le rappel, la distance de `Jaccard`, `BLEU`, etc., ne permettent pas de quantifier la qualité d'un classement. Nous avons défini une procédure d'évaluation pour comparer la qualité des classements établis par la méthode lexicographique et la méthode par surclassement. Cette évaluation porte évidemment sur les 157 requêtes annotées par l'expert. Nous avons utilisé une mesure d'évaluation par pénalité pour comparer les classements établis par les deux méthodes avec les annotations de l'expert. Au cours de cette évaluation, nous comptabilisons le nombre de requêtes pour lesquelles une rubrique annotée comme correcte se retrouve en tête de classement et désignons cette valeur sous le terme de $correctesRemontees$. Puis, afin d'évaluer le reste du classement, nous déterminons le score comptabilisant les erreurs (noté $nbErreursClassement$) commises lors de la comparaison des rubriques concurrentes. Si une rubrique annotée comme incorrecte est mieux classée qu'une rubrique correcte, une pénalité de classement d'une valeur de 2 est infligée à la méthode. Si une rubrique incorrecte est mieux classée que N rubriques correctes, alors la pénalité devient $2 * N$. Dans le cas où une rubrique annotée comme indifférente se trouve mieux classée qu'une hypothèse correcte alors la pénalité infligée est de 1.

Il est évident que le corpus de référence n'est pas suffisamment volumineux pour observer statistiquement des régularités. Nous n'avons donc pas pu exploiter les différentes heuristiques statistiques proposées lors de la section 4.3.3 pour guider la construction d'un modèle de préférences adapté à ce nouveau cas de contrôle. Avec l'aide d'un linguiste expert du domaine concerné, nous avons affecté aux différents paramètres décisionnels des valeurs nous permettant d'espérer une meilleure utilisation des 4 critères disponibles. Après une explication de la sémantique as-

[13] Nous précisions ici que le membre de l'équipe en charge de ce projet depuis sa création a constitué pour nous l'expert de ce domaine qui est l'indexation de requêtes soumises à un annuaire de professionnels

sociée à chaque paramètre (poids, seuil de préférence, indifférence, veto), un modèle visant à exploiter au maximum la complémentarité des critères a rapidement été mis en place. Comme en témoigne le tableau 5.1, les poids associés aux différents critères reprennent l'ordre d'importance initialement établi. Afin de nuancer l'impact d'une très faible supériorité d'une rubrique par rapport à une autre, des seuils de préférence et d'indifférence ont été définis sur les différents critères.

TAB. 5.1 – Modèle de préférences établi par l'expert pour le classement des rubriques métiers concurrentes

`Critères`	`Poids`	`Seuil préf.`	`Seuil indiff.`	`Seuil veto`
score	0.4	1	0	3
redondance	0.2	2	1	_
rang	0.2	4	2	100
homogénéité	0.2	2	1	_

5.1.4 Résultats et interprétations

Le tableau 5.2 illustre les résultats obtenus par les deux stratégies de classement en appliquant la méthode d'évaluation présentée précédemment.

TAB. 5.2 – Résultats de l'évaluation des stratégies de classement

méthode	`lexicographique`	`surclassement`
$correctesRemontees$	139	141
$nbErreursClassement$	329	308

On observe que l'usage d'une méthode par surclassement n'entraîne pas une amélioration significative des résultats sur notre ensemble de test. D'une manière positive, on constate tout d'abord que la prise en compte de plusieurs critères n'entraîne pas de régression, et pour deux requêtes, une analyse correcte est remontée en tête de classement à l'aide de notre approche par surclassement. De même la pénalité globale infligée sur les classements construits diminue de 21.

Bien que ce premier cas d'expérimentation ne mette pas en avant l'efficacité des approches par surclassement et de la prise en compte de connaissances expertes, il nous apporte tout de même des informations intéressantes sur la méthode et le processus contrôlé.

Une étude de la variation des performances obtenues par les rubriques concurrentes sur les différents critères nous a permis de constater que les critères ne semblent pas complémentaires et surtout très peu discriminants. En effet, pour 80% des requêtes, les performances obtenues par les rubriques concurrentes sur les critères de score, d'homogénéité et de redondance sont identiques et donc non discriminantes. Le critère de rang apparaît donc comme un critère prépondérant. Ainsi, en n'exploitant que le critère de rang, les résultats obtenus ne sont pas significativement moins bons ($correctesRemontees = 119$ et $nbErreursClassement = 410$).

Ces résultats s'expliquent également par la simplicité des requêtes du corpus de référence utilisé. En effet, en exploitant en priorité le corpus des 100 requêtes les plus fréquemment soumises

au service d'annuaire, nous nous sommes restreint aux requêtes les plus simples, composées la plupart du temps d'un ou de deux mots (en moyenne 1,9 mots). De plus, on constate que la faible amélioration des résultats obtenue par l'application de notre approche de contrôle par surclassement concerne des requêtes complexes comme "vente aux enchères publiques : société de vente volontaire de meubles" ou "salons Fabio Salsa ouvert en nocturne le jeudi". Afin d'attester l'intérêt d'une approche multicritère de contrôle et plus précisément l'apport des méthodes par surclassement, il nous faudrait nous focaliser sur le classement des rubriques concurrentes générées sur ce genre de requêtes complexes.

De plus, le processus d'indexation réalisé par `TiLT` et plus particulièrement les ressources linguistiques qui lui sont associées ont été développés afin de garantir des résultats pertinents pour les requêtes les plus fréquemment soumises au service d'annuaire de professionnels. Ainsi, le comportement du processus d'indexation a été fortement guidé pour que ces requêtes récurrentes soient correctement traitées. C'est sur ces mêmes requêtes que nous avons effectué notre évaluation. Il s'agit de la raison principale expliquant la faible progression obtenue sur ce jeu de test très spécifique. Ceci confirme donc qu'il serait intéressant de désormais procéder à une évaluation des stratégies de contrôles sur des requêtes plus complexes, mais le risque de dégrader les résultats obtenus sur les requêtes les plus fréquentes n'est évidemment pas souhaitable et ne correspondrait pas aux attentes d'une telle application commerciale.

Cependant, d'un point de vue expérimental, il serait intéressant de compléter cette comparaison des deux méthodes de classement sur des requêtes plus complexes et sur lesquelles les limites de la méthode lexicographique pourraient être mises en évidence.

Nous avons tout de même pu constater la facilité d'utilisation du module de contrôle `Beslissing`. Les quatre informations distinctives associées aux différentes rubriques générées ont été tout d'abord déclarées en tant que critères stockés de manière centralisée dans le module de contrôle décisionnel, sur lesquels un opérateur de surclassement a été appliqué pour construire un classement des rubriques concurrentes. Bien que ces travaux n'aient pas pu être entièrement exploités, cette normalisation du classement des rubriques concurrentes en tant que problème décisionnel permet d'expérimenter plusieurs stratégies de contrôle (variation des modèles de préférences) sans avoir à modifier le code de `TiLT` et le recompiler. Ainsi, sur un corpus de requêtes plus complexes, l'externalisation des configurations décisionnelles nous permettra de construire progressivement une stratégie de contrôle exploitant au mieux la connaissance apportée par chacun des critères.

De plus, l'usage de notre module de contrôle permet une meilleure traçabilité des éléments décisionnels disponibles et ainsi de comprendre les raisons de la construction d'un classement. Ceci nous a notamment permis de constater que pour la plupart des requêtes de notre corpus, seul le critère de rang permettait de différencier les rubriques concurrentes.

Conclusion d'expérimentation

Le processus d'indexation réalisé par `TiLT` pour un service d'annuaire de professionnels constitue un cas typique d'indétermination. Nous avons en effet constaté que de nombreuses rubriques métiers concurrentes sont générées suite à l'analyse d'une requête utilisateur. Face à la nécessité de disposer d'une évaluation de la pertinence relative des différentes rubriques, des connaissances supplémentaires ont été intégrées en tant que critères de comparaison afin notamment d'établir un classement de ces hypothèses concurrentes.

Ces critères ont été initialement exploités à l'aide d'une méthode lexicographique, et nous avons cherché à travers cette expérimentation à obtenir une meilleure exploitation de l'information apportée par chacun d'eux à travers l'usage de notre module de contrôle par surclassement. Nous avons constaté que la conception du module de contrôle `Beslissing` permettait facilement d'externaliser la gestion du processus de classement et ainsi de permettre à un expert d'exprimer ses connaissances à l'aide des différents paramètres préférentiels. L'application du module de contrôle à ce premier cas d'indétermination a confirmé la généricité de nos travaux.

Nous avons cependant constaté dans ce cas précis et sur ces requêtes particulières que l'usage d'une approche par surclassement ne conduisait pas à de meilleurs résultats. Ceci s'explique par la particularité du contexte d'évaluation. En effet, nous avons utilisé un corpus composé principalement de requêtes simples et très fréquentes pour lesquelles l'apport d'un contrôle multicritère n'est pas significatif, car les ressources linguistiques associées au processus d'indexation ont été développées en vue de donner des résultats pertinents pour les requêtes les plus fréquentes.

On pourrait en déduire que la complexité induite par l'application d'une approche par surclassement pour le classement des rubriques concurrentes ainsi que la faible progression obtenue justifient l'usage d'une méthode simple basée sur un ordre lexicographique. Cependant, les bons résultats obtenus par cette dernière méthode s'expliquent par le fait que sur les requêtes utilisées au cours de l'évaluation, l'approche multicritère n'était pas pertinente. En effet, nous avons observé que seul le critère de rang était réellement discriminant et donc à prendre en compte.

D'un point de vue expérimental, mais allant également à l'encontre des intérêts commerciaux du domaine applicatif, il serait intéressant de poursuivre cette comparaison des méthodes de classement sur des cas plus complexes pour lesquels l'apport des autres critères de comparaison pourrait être mis en évidence. Cette perspective permettrait d'avoir une meilleure évaluation de notre module de contrôle pour les problématiques de classement. Cette expérimentation nous a montré l'importance du choix des critères de comparaison à exploiter lors d'une stratégie de contrôle. L'efficacité d'une stratégie de contrôle repose autant sur la qualité de la méthode exploitée que sur la pertinence de la famille de critères construite [Roy and Bouyssou, 1993, chap.2].

5.2 Identification des cas de reprise anaphorique : problématique de tri

Ces travaux s'inscrivent dans le cadre d'une collaboration avec un doctorant, Olivier Tardif, travaillant au sein de `Orange Labs` sur la problématique de l'identification des reprises anaphoriques dans un texte. Cette expérimentation vise à montrer en quoi notre module de contrôle est applicable pour la résolution de ce problème. Comme le lecteur pourra le constater, notre intervention dans cette tâche nous place en tant qu'homme d'étude chargé de la mise en place d'une stratégie décisionnelle de contrôle, où Olivier Tardiff apparaît comme expert du domaine.

5.2.1 Présentation du problème

La résolution des coréférences constitue une étape intermédiaire dans un processus de TALN, dont l'objectif est d'identifier les paires d'expressions qui partagent un lien de coréférence. Deux expressions sont considérées comme coréférentes si elles réfèrent à un même objet, une même entité du monde. Les expressions qui forment ces paires peuvent être de plusieurs types :

1. expression nominale-ANTÉCÉDENT / expression nominale-REPRISE :
   Ex. : "L'enfant-ANTÉCÉDENT agé de 4 ans est (...). Cette jeune personne-REPRISE (...)"
2. expression nominale-ANTÉCÉDENT / expression pronominale-REPRISE :
   Ex. : "Prenez 5kg de malt-ANTÉCÉDENT, concassez le-REPRISE"
3. expression pronominale-ANTÉCÉDENT / expression pronominale-REPRISE :
   Ex. : "Il-ANTÉCÉDENT est là. Attrapez le-REPRISE !"
4. entité nommée-ANTÉCÉDENT / entité nommée-REPRISE :
   Ex. : "Nouvelle victoire de Mickael Schumacher-ANTÉCÉDENT. Mais qui pourra battre Schumi-REPRISE ?"
5. entité nommée-ANTÉCÉDENT / expression pronominale-REPRISE :
   Ex. : "Boris Becker-ANTÉCÉDENT a battu (...). Il-REPRISE accède donc à la finale (...)."
6. entité nommée-ANTÉCÉDENT / expression nominale-REPRISE :
   Ex. : "Mikail Gorbatchev-ANTÉCÉDENT prononcera un discours lundi (...). Le dirigeant soviétique-REPRISE devrait annoncer (...). "

L'identification de ce phénomène complexe et varié constitue un des exercices les plus délicats dans le domaine du TALN. Cette tâche n'en reste pas moins indispensable pour de nombreux autres modules de traitement tels que la traduction automatique, la construction automatique de résumés, etc.

Comme nous l'avons signalé précédemment, cette expérimentation s'appuie sur des travaux visant à compléter la chaîne de traitement TiLT par un module d'identification des paires d'expressions coréférentes. Dans un premier temps, cette étude s'est focalisée uniquement sur l'identification des paires coréférentes dont l'antécédent est une entité nommée, c'est-à-dire les trois dernières configurations présentées lors de l'énumération précédente des différents types de coréférence. Cette restriction du problème est motivée par la volonté de s'attaquer aux coréférences les plus facilement identifiables et l'intuition que la résolution de ces cas précis fournira des informations intéressantes pour l'identification des chaînes complètes de coréférence. Pour plus de détails sur cette sous-tâche de la résolution de la coréférence, le lecteur intéressé pourra consulter les communications liées à la campagne d'évaluation des systèmes de résolution de la coréférence MUC-7 [Chinchor and Hirschmann, 1997] dont le cadre et les objectifs étaient comparables.

Lors de cette seconde section, nous allons constater que les travaux menés ont permis de définir un cas typique d'indétermination sur lequel notre méthodologie décisionnelle de contrôle peut être appliquée. Avant de nous intéresser à l'application de notre méthode de contrôle, nous allons voir que les travaux menés lors de cette étude de la coréférence par Olivier Tardif ont conduit au développement d'un algorithme d'extraction de paires d'expressions candidates et à l'identification des informations à exploiter pour attester de la validité d'un candidat, c'est-à-dire statuer sur un lien de coréférence entre deux expressions. Nous verrons également qu'un corpus de référence a été établi et qu'il offre des possibilités intéressantes d'apprentissage et d'évaluation.

Extraction des paires candidates

Il est évident que quelle que soit la démarche envisagée, la résolution des cas de coréférence repose en premier lieu sur une identification des expressions pouvant potentiellement avoir un rôle d'antécédent ou de reprise. Nous venons de voir que les types d'expressions envisagées correspondent soit à des groupes nominaux ou pronominaux[14], soit à des entités nommées. L'obtention de ces expressions repose alors sur une analyse syntaxique de surface permettant d'effectuer un découpage et une catégorisation des différents constituants syntaxiques des phrases traitées (groupes nominaux, verbaux, pronominaux, adverbiaux, prépositionnels, etc.) et sur une étape plus spécifique d'identification des entités nommées. La figure 5.3 illustre le processus d'analyse linguistique effectué par `TiLT` pour fournir les informations, notamment structurelles, nécessaires à l'application d'une procédure dédiée à la coréférence.

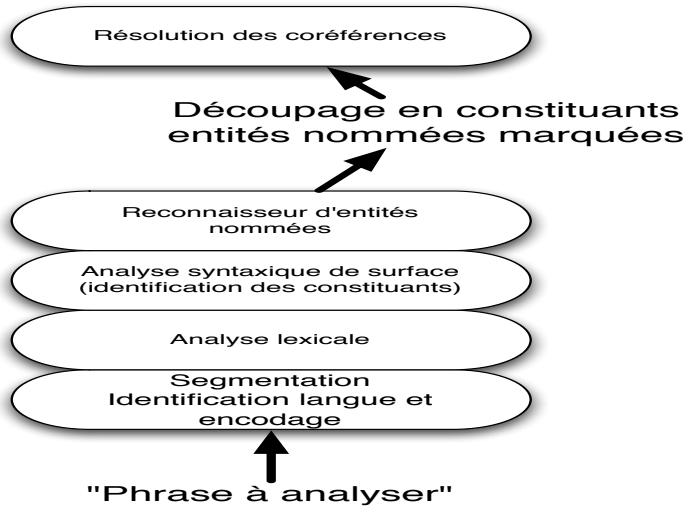

FIG. 5.3 – Utilisation de la chaîne de traitement `TiLT` pour la résolution des chaînes de coréférence

Les informations générées par le processus d'analyse linguistique présenté précédemment sont ensuite utilisées pour effectuer la première étape de la procédure de résolution des coréférences, à savoir l'extraction des paires d'expressions candidates. L'algorithme d'extraction suit une démarche relativement simple, puisqu'il parcourt séquentiellement le texte et considère chaque entité nommée rencontrée comme un antécédent potentiel, et construit des paires d'expressions candidates avec toutes les autres expressions (nominale, pronominale ou entité nommée) qui lui succèdent[15], en se limitant à une fenêtre d'analyse donnée (3 paragraphes).

Cette première phase d'extraction permet de constituer un ensemble de paires d'expressions candidates, qui doivent désormais être individuellement évaluées, ce qui permettra de décider si un lien de coréférence existe entre les deux expressions.

[14] Bien que la notion de groupe soit ici peu pertinente dans la mesure où ce groupe est la plupart du temps réduit à un pronom.

[15] Le cas des cataphores, où la reprise précède l'antécédent, n'est donc pas géré pour le moment.

Informations pour établir les liens de coréférence entre les expressions

De multiples informations de différentes natures permettent d'identifier un cas de coréférence entre deux expressions. Un travail bibliographique et une étude sur corpus (voir section 5.2.1) ont permis à Olivier Tardif [Tardif, 2005] d'identifier les attributs les plus pertinents à utiliser, mais également déterminer ceux qui sont accessibles et disponibles dans notre contexte d'analyse. En effet, l'identification de nombreux cas de coréférence nécessite l'usage de connaissances qualifiées d'universelles ou encyclopédiques, mais qui sont cependant rarement disponibles, formalisables et exploitables électroniquement. Il est donc nécessaire de se focaliser sur des aspects dits de surface ou linguistiques.

Les informations exploitables dans ce que nous avons nommé le contexte d'analyse sont donc composées de connaissances linguistiques générées par `TiLT` (figure 5.3) et de données statistiques ou de mesures directement extraites du texte analysé. L'identification des cas de coréférence parmi les paires extraites repose donc sur l'exploitation de différents attributs qui qualifient soit la paire d'expressions, soit chaque expression individuellement. Ces informations sont comme nous venons de le signaler de différentes natures (l'annexe D.2 propose une énumération des 25 informations prises en compte) :

- typographique (similarité entre les deux expressions, la reprise est une sous-chaîne de l'antécédent, le pourcentage de mots en commun, etc.)
- catégorielle / syntaxique (l'antécédent est un sujet de la phrase, les deux expressions ont la même fonction syntaxique, possède une marque de défini, d'indéfini, de possessif, accord genre-nombre, etc.)
- statistique ou contextuelle : (distance en mots, expressions et phrases entre les expressions, fréquence de l'antécédent dans le corpus, l'antécédent est l'entité nommée la plus proche de la reprise potentielle, etc.)

La validation ou la non validation d'une paire d'expressions candidate en tant que cas de coréférence repose donc sur l'interprétation de ces attributs, qui sont soit générés par `TiLT`, soit extraits de sources de connaissances supplémentaires (thésaurus par exemple), soit calculés à partir du texte analysé.

Un corpus de référence disponible pour l'apprentissage et l'évaluation

Lors de son étude du phénomène de la coréférence, Olivier Tardif a également procédé à la construction d'un corpus de référence afin d'ouvrir des perspectives vers l'exploitation de méthodes issues de l'apprentissage automatique, comme il est d'usage sur cette problématique.

Ce sous-corpus est constitué de 80 articles[16] du journal `Le Monde` datant de 1989 et 1990. Ces articles, chacun d'une longueur moyenne de 500 à 1000 mots, appartiennent à des domaines variés de l'actualité mais dans un style journalistique identique.

L'intégralité du corpus a dans un premier temps été analysée par `TiLT` en appliquant le processus d'analyse illustré par la figure 5.3. Cette analyse avait pour but de fournir notamment les informations syntaxiques nécessaires à l'extraction des expressions formant ensuite les paires candidates, mais également les informations nécessaires à l'identification des cas de coréférence (section 5.2.1). Les analyses (regroupement en constituant, catégories morpho-syntaxiques, etc.) générées par `TiLT` ont été manuellement vérifiées et corrigées en cas d'erreur. Afin de constituer

[16] Extraits d'un corpus complet (Le Monde 89-90) composé de 8000 textes.

un corpus de référence à la fois pour ces informations linguistiques et pour le phénomène de la coréférence, les paires d'expressions partageant un lien de coréférence ont été identifiées et catégorisées (type des expressions), constituant au total un ensemble de 3504 paires d'expressions coréférentes.

Un alignement entre les paires d'expressions générées par l'algorithme d'extraction et le corpus de référence nous a permis de construire trois tables de performances, une pour chaque configuration de types de paires. Nous rappelons que ces tables de performances sont constituées des hypothèses, des attributs qui les qualifient et de leur classe d'appartenance. Comme nous allons le voir par la suite, les deux seules classes considérées sont la classe des paires coréférentes (C_1) et la classe des paires non coréférentes (C_0). Le tableau 5.3 illustre sur quelques exemples le contenu des tables de performances dédiées respectivement aux paires entité nommée-entité nommée (notées par la suite `NPR-NPR`), entité nommée-expression nominale (notées par la suite `NPR-NCOM`) et entité nommée-expression pronominale (notées par la suite `NPR-PRON`).

En disposant d'un corpus de référence contenant à la fois la validation des paires coréférentes et des informations qui les qualifient, il devient possible d'évaluer une méthodologie de résolution des cas de coréférence sachant un ensemble de critères donné. Un extrait de ce corpus est proposé en annexe D.1.

TAB. 5.3 – Extraits des tables de performances par un alignement entre les paires candidates extraites du texte et le corpus de référence. La description des attributs et de leur type est proposée en annexe D.2.

Paires d'expressions		**Attributs-critères**					**Classe**
antécédent	reprise	isClosest	countOcc	...	subString	similModif	C_1 / C_0
NPR-NPR							
"Hamed Karoui"	"Karoui"	0	3	...	1	0	C_1
"Mr Karoui"	"Mr Baccouche"	1	2	...	0	1	C_0
NPR-NCOM							
"Mr H. Karoui"	"le ministre"	0	1	...	0	0	C_1
"les États-Unis"	"le prix nobel"	1	4	...	0	0	C_0
NPR-PRON							
"Hedi Baccouche"	"lui"	0	5	...	0	0	C_1
"Mr Bourguiba"	"elle"	1	1	...	0	0	C_0

Une problématique décisionnelle de tri

L'ensemble de ces travaux réalisés dans le domaine de la résolution de la coréférence fournit un cadre d'expérimentation pour l'application de notre méthodologie de contrôle. En effet, les paires d'expressions extraites constituent l'ensemble des hypothèses concurrentes H à contrôler. Les attributs qui les qualifient correspondent à un ensemble de critères G de type numérique ou binaire, permettant d'évaluer selon différents axes la validité des différentes hypothèses candidates.

Comme nous l'avons souligné dans la section consacrée à la formalisation d'un problème d'AMCD (section 3.1.2), les critères utilisés lors de la construction des relations de surclassement doivent correspondre à des fonctions croissantes, c'est-à-dire que plus la performance obtenue par une hypothèse est forte, plus la validité de cette hypothèse est attestée sur le critère concerné. Bien que la plupart des critères utilisés (annexe D.2) suive intrinsèquement une telle variation croissante, certains critères ont dû être modifiés. C'est notamment le cas de certaines mesures, critères numériques, ou de certaines propriétés, critères binaires, servant à identifier des marques linguistiques sur une expression qui sont incompatibles avec un rôle de reprise anaphorique. Parmi ces critères dont le sens de variation a été inversé, nous pouvons par exemple citer la présence dans la reprise candidate d'une marque indéfinie (pronom ou article), le fait que deux entités nommées apparaissent dans une apposition, que les deux expressions soient dans la même structure actancielle d'un verbe, l'absence de déterminant dans une expression nominale, etc. Cette transformation en fonctions croissantes fut cependant un peu plus délicate pour les critères mesurant des distances entre les deux expressions d'une paire. Il nous est apparu *a priori* indispensable d'exploiter différemment les mesures de distance en fonction du type des expressions comparées. En effet, il semble qu'une expression pronominale ou nominale se situe en général dans un contexte proche de son antécédent, mais que cette proximité ne soit pas forcément pertinente ou discriminante lorsqu'il s'agit d'une paire d'entités nommées coréférentes. Nous avons donc inversé la variation des mesures de distance entre les expressions (plus elles sont proches plus la performance est forte), bien que cette information perde alors son sens pour les paires candidates composées de deux entités nommées. Nous verrons cependant que cette modification n'a pas d'impact sur la procédure de résolution dans la mesure où les critères de distance ne sont pas vraiment pertinents pour l'identification des liens de coréférence entre des entités nommées. Ainsi, pour ces critères représentant les propriétés que ne doivent pas posséder une paire d'expressions pour être éventuellement considérée comme une coréférence, nous avons transformé le sens de variation des performances en les remplaçant par leur inverse ($\frac{1}{performance}$).

On constate donc que l'exploitation des résultats de l'algorithme d'extraction des paires d'expressions potentiellement coréférentes constitue un cas typique d'indétermination, où l'objectif est de discriminer l'ensemble de ces candidats en deux sous-ensembles : d'une part les paires d'expressions pour lesquelles un lien de coréférence peut être établi et d'autre part les paires candidates ne correspondant pas à un cas de coréférence. Cette tâche de résolution a donc été fréquemment considérée comme un problème de classification [Weissenbacher and Nazarenko, 2007b] et [Tardif, 2006]. Mais la transformation des critères d'identification des cas de coréférences en tant que fonctions strictement croissantes nous permet désormais de considérer cette tâche comme un problème de tri comprenant deux classes : une classe "haute" des paires validées comme coréférentes (classe C_1) et une classe "basse" des paires non validées (classe C_0).

Outre le fait que cette tâche constitue un cadre d'évaluation de notre méthodologie de contrôle, tout au moins pour la problématique de tri, ce travail émane également d'un constat de manque de transparence et de flexibilité des méthodes principalement utilisées pour la résolution de la coréférence. En effet, dans le cadre d'une première tentative d'exploitation du corpus de référence créé, un classifieur bayésien[17] a été utilisé afin d'estimer le taux de précision envisageable pour cette problématique à l'aide du corpus disponible [Tardif, 2006]. Bien que les résultats obtenus soient corrects en terme de précision (environ 0,6 en moyenne pour les trois types de paires traitées), leur interprétabilité reste obscure, tout comme la plupart des résultats

[17] Fourni par l'`API WEKA` - `http://weka.sourceforge.net`

obtenus par des méthodes purement statistiques. En effet, le processus de classification effectué par ce genre de méthodes reste difficilement interprétable et surtout ne laisse aucune place à un expert du domaine pour l'intégration de ses connaissances et intuitions sur le phénomène étudié. Le processus d'identification des coréférences et les résultats générés sont donc essentiellement dépendants du corpus d'apprentissage sur lequel la méthode a été entraînée. Nous allons voir que dans certaines conditions, notamment pour l'identification de phénomènes minoritaires pourtant discriminants, ces méthodes sont peu efficaces.

À travers ce cas d'expérimentation, nous cherchons donc à montrer que notre méthodologie de contrôle décisionnel par surclassement est exploitable pour un tel cas réel d'indétermination, et qu'elle offre également des perspectives intéressantes et novatrices en permettant notamment à un expert humain de maîtriser ou d'influencer le processus d'identification des coréférences à l'aide de ses propres intuitions et connaissances.

5.2.2 Présentation de la démarche expérimentale

Nous venons de constater que suite à l'application d'un algorithme d'extraction des paires d'expressions candidates, le processus de résolution des coréférences se trouvait dans une situation que nous avons précédemment qualifiée de "point de décision" (section 2.1.3). Ce "point de décision" est donc caractérisé par la présence d'un ensemble d'hypothèses candidates et la disponibilité d'une famille de critères permettant d'évaluer selon différents aspects leur validité respective. Attester de la validité d'une hypothèse, c'est-à-dire confirmer l'existence d'un lien de coréférence entre les deux expressions, repose sur la prise en compte de ces différents critères hétérogènes et potentiellement contradictoires. Nous avons également spécifié précédemment que l'ensemble de ces fonctions critères (voir section 3.1.2) respectait désormais une variation strictement croissante, nous permettant d'appréhender ce problème sous forme d'un tri basé sur deux classes : celle des hypothèses validées et celle des hypothèses non validées.

Lors de la présentation des méthodes d'AMCD par surclassement, nous avons souligné que la validité des décisions émises reposait à la fois sur la pertinence des critères utilisés, et sur l'adéquation, vis-à-vis du contexte décisionnel concerné, d'un ensemble de paramètres décisionnels formant un modèle de préférences. L'étude effectuée par un expert sur le phénomène de la coréférence a déjà permis de mettre en avant un ensemble de critères jugés adéquats pour l'identification de ces liens entre expressions. Notre expérimentation s'intéresse donc plus particulièrement aux processus de mise en œuvre des modèles de préférences, qui définissent la façon dont ces critères doivent être interprétés. Nous comparerons ainsi dans un premier temps les différentes approches envisagées pour la construction de ces modèles de préférences. Nous verrons ainsi que cet ensemble de paramètres décisionnels permet à un expert d'intégrer ses intuitions et connaissances sur la manière dont la procédure d'identification des cas de coréférence doit opérer. Nous verrons ensuite que les valeurs suggérées par les heuristiques que nous avons définies dans la section 4.3.3 sont pertinentes, mais surtout qu'elles permettent à un expert d'identifier plus facilement les critères utiles et la façon dont ils doivent être utilisés.

À partir de ces premiers constats, nous pourrons dans un second temps revenir sur l'intérêt même des approches par surclassement, c'est-à-dire la prise en compte de tous ces paramètres décisionnels.

Nous avons également signalé précédemment (section 5.2.1) que trois tables de performances, une pour chaque type de paires (NPR-NPR, NPR-NCOM, NPR-PRON), avaient été construites à partir d'un alignement entre le corpus de référence et les paires d'expressions extraites de ce même corpus par l'algorithme d'extraction. Ces tables de performances ont été divisées en deux parties, l'une constituant un ensemble d'exemples de références dit d'apprentissage, c'est-à-dire dédié à l'application des heuristiques statistiques proposées dans la section 4.3.3, et l'autre constituant un ensemble d'exemples différents pour l'évaluation des différentes stratégies de contrôle. Ce découpage a été effectué selon un rapport arbitraire d'environ 62% des tables initiales pour l'apprentissage et 38% pour l'évaluation.

Le tableau 5.4 présente les dimensions des différentes tables de performances créées en séparant les tables utilisées pour l'apprentissage et l'évaluation. Pour chacun de ces deux usages, nous séparons également les comptages en fonction du type des paires : NPR-NPR, NPR-NCOM ou NPR-PRON.

Tab. 5.4 – Répartition des paires d'expressions extraites dans les tables de performances

Type	Classe	NPR-NPR	NPR-NCOM	NPR-PRON
Apprentissage	Non valides	4653	5891	4829
	Valides	805	354	322
Évaluation	Non valides	2981	3422	3085
	Valides	244	237	217
Total	Non Valides	7634	9313	7924
	Valides	1049	591	539
	Extraites	8683	9904	8463

Dans notre contexte d'évaluation, la problématique de résolution des liens de coréférences repose donc sur l'identification des paires d'expressions que nous avons qualifiées de valides, c'est-à-dire partageant une référence vers une entité commune, en opposition aux paires d'expressions non valides extraites du texte, pour lesquelles aucun lien de coréférence existe. On constate que l'une des difficultés de cet exercice provient de la très faible proportion de paires positives par rapport au nombre total de paires extraites (environ 8,9% sur les tables d'apprentissage et 7,4% sur les tables d'évaluation).

Face à cette problématique de classification ou de tri, de nombreuses méthodes semblent applicables. Dans ces travaux visant à proposer un module de résolution des liens de coréférence, la première idée d'Olivier Tardif était d'appliquer sur des données représentatives, des algorithmes de classification et plus précisément un classifieur bayésien, en utilisant une des plateformes logicielles proposant ces outils "clés en main", comme WEKA[18] ou YALE [Mierswa *et al.*, 2003]. En effet, les tables de performances d'apprentissage peuvent aisément être formalisées afin d'être conformes au format d'entrée attendu par ces outils. Cependant, constatant qu'il était délicat d'interpréter les résultats générés par ces outils et impossible d'influencer le processus d'apprentissage à l'aide de connaissances empiriques, ces expérimentations autour des outils d'apprentissage n'ont pas été poursuivies. En effet, Olivier Tardif, en tant qu'expert des aspects linguistiques du phénomène étudié et non des méthodes d'apprentissage, a rapidement constaté l'absence de lisibilité et d'interprétabilité des résultats obtenus, ce qui ne lui permet-

[18] http://weka.sourceforge.net

tait pas de valider ou de compléter ses connaissances et intuitions sur la résolution des liens de coréférence. Ce constat explique donc le fait que nous nous soyons intéressé à l'apport d'une approche par surclassement sur ce problème, qui a l'avantage de prendre en compte l'expert lors du processus décisionnel de tri.

Évaluation sur différents modèles de préférences : expert, observé et mixte

Afin d'évaluer notre approche de contrôle par surclassement des processus de TALN, et dans ce cas particulier pour une problématique de tri, nous avons cherché à définir trois modèles de préférences (un pour chaque type de paire) traduisant les intuitions formulées par un expert du domaine, en l'occurrence Olivier Tardif. Ainsi, après une explication de la sémantique associée à l'ensemble des paramètres décisionnels (poids, seuils de préférence, indifférence, veto et limites d'acceptabilité) qui composent un modèle de préférences, nous avons sollicité l'expert pour qu'il détermine des valeurs traduisant au mieux ses intuitions sur la façon dont les critères doivent être exploités.

Nous lui avons tout d'abord demandé d'établir, pour chacun des types de paires d'expressions envisagés, un classement des critères selon leur importance, c'est-à-dire leur capacité à discriminer des paires coréférentes par rapport à des paires non coréférentes. Puis, nous lui avons demandé d'associer à cette hiérarchie une distribution de poids numériques sur une échelle de 0 à 10, qui reflète au mieux ses intuitions sur l'importance relative des critères. Il est à noter que le fait d'affecter un poids nul à un critère revient à l'écarter de l'ensemble des critères pris en compte lors de la construction des relations de surclassement. Dans un second temps, nous lui avons demandé de déterminer des limites d'acceptabilité pour chacun des critères conservés (poids supérieur à 0), en rappelant que ces limites constituaient des seuils d'acceptabilité pour les performances de critères. Finalement, nous lui avons donné la possibilité de déterminer des marges de tolérance ou d'intolérance pour chacune de ces limites, que nous avons par la suite traduites en seuils de préférence, indifférence et de veto. Cette démarche de capture des intuitions de l'expert nous a permis d'obtenir des modèles de préférences qualifiés d'experts. Les tableaux 5.5, 5.6 et 5.7 présentent les différents paramètres définis par l'expert pour les cinq critères qu'il a identifiés comme prépondérants (la totalité des modèles de préférences experts est proposée en annexe D.3.1).

Tab. 5.5 – Paramètres préférentiels associés par l'expert pour les paires `NPR-NPR` : extrait du modèle de préférences complet proposé en annexe D.3.1

NPR-NPR

Rang	**Critere**	**Poids** $[0, 10]$	**Seuil pref.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	strSimil	4	0,2	0,1	0,27	0,3
2	subString	3	-	-	-	0,5
3	wordsInCommon	1	0,1	0,05	0,25	0,15
4	isAcronym	1	-	-	-	0,2
5	countOcc	0,5	0,02	0,04	-	0,1
...	...	...	...	...	...	...

TAB. 5.6 – Paramètres préférentiels associés par l'expert pour les paires `NPR-NCOM` : extrait du modèle de préférences complet proposé en annexe D.3.1

NPR-NCOM

Rang	**Critere**	**Poids** $[0, 10]$	**Seuil pref.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	countOcc	1,5	-	-	-	0.05
2	appo	1,5	-	-	-	0,5
3	isClosestCandidate	1	-	-		0,5
4	distExp	1	0,15	0,05	-	0,55
5	distSent	1	0,1	0,05	-	0,65
...	...	...	...	...	...	...

TAB. 5.7 – Paramètres préférentiels associés par l'expert pour les paires `NPR-PRON` : extrait du modèle de préférences complet proposé en annexe D.3.1

NPR-PRON

Rang	**Critere**	**Poids** $[0, 10]$	**Seuil pref.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	isClosestCandidate	2	0,2	0,1	-	0,2
2	countOcc	2	0,2	0,005	-	0,05
3	distExp	1	0,15	0,1	0,5	0,7
4	distSent	1	0,1	0,05	0,3	0,85
5	distTer	1	0,15	0,1	0,5	0,7
...	...	...	...	...	...	...

Ces modèles de préférences expert ont ensuite été évalués sur les sous-tables de performances dédiées à l'évaluation. Le tableau 5.8 présente les résultats obtenus par l'application de notre méthode de contrôle par tri, en s'appuyant sur les modèles de préférences experts. L'interprétation de ces résultats est proposée plus loin dans la section 5.2.3, afin de pouvoir notamment comparer les différentes stratégies de tri évaluées (modèles de préférences experts, suggérés et mixtes).

TAB. 5.8 – Résultats obtenus en utilisant les modèles de préférences définis *a priori* par un expert

	`NPR-NPR`	`NPR-NCOM`	`NPR-PRON`
Précision	0,87	0,89	0,93
Rappel	0,94	0,32	0,42
F-mesure	0,90	0,47	0,58

Lors de cette démarche de construction d'un modèle de préférences respectant au mieux les connaissances et les intuitions d'un expert, nous avons constaté que la construction d'un classement des critères selon leur importance relative, ainsi que la proposition d'une distribution de poids associées à ce classement, ne posait pas de réelles difficultés. Cependant, l'identification de valeurs possibles et pertinentes pour les autres paramètres tels que les seuils de préférence, indifférence, veto est apparue comme réellement plus délicate. Cette difficulté s'explique principalement par le fait que l'utilité de ces différents paramètres n'est perceptible que pour les personnes déjà

expérimentées dans l'utilisation des méthodes par surclassement. Ce n'est qu'après avoir appliqué plusieurs stratégies décisionnelles que l'on comprend le réel apport de ces informations.

Les différentes heuristiques que nous proposons (section 4.3.3) visent justement à pallier cette difficulté. Il est important de rappeler que ces heuristiques n'ont pas pour vocation d'inférer un modèle de préférences "optimal" à partir d'une table de performances, mais bien de constituer des outils d'aide à l'élaboration d'un modèle de préférences au service d'un expert. Afin d'évaluer la pertinence et la validité de ces heuristiques, nous avons construit des modèles de préférences à partir uniquement d'observations statistiques effectuées sur les sous-tables de performances dédiées à l'apprentissage. L'application de l'algorithme `RELIEF` pour l'obtention d'une distribution de poids et des heuristiques pour les autres paramètres décisionnels (seuils de préférence, d'indifférence, veto et coupe) nous a conduit aux modèles de préférences présentés partiellement par les tableaux 5.9, 5.10 et 5.11, et de manière plus complète en annexe D.3.2. Nous désignons par la suite ces modèles comme des modèles de préférences suggérés. Les figures D.1, D.2 et D.3 de l'annexe D.3.2.0 illustrent trois exemples de courbes de distributions exploitées pour suggérer des valeurs aux différents paramètres préférentiels.

Tab. 5.9 – Paramètres préférentiels suggérés à partir des tables de performances dédiées à l'apprentissage pour les paires `NPR-NPR` : extrait du modèle de préférences complet proposé en annexe D.3.2

NPR-NPR

Rang	**Critere**	**Poids** $[0,1]$	**Seuil pref.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	wordsInCommon	0,25	0,2	0,0	0,0005	0,21
2	subString	0,25	-	-	-	1
3	isSimil	0,2	-	-	-	1
4	strSimil	0,2	0,26	0,12	0,54	0,71
5	countOcc	0,03	2	1	-	2
...	...	...	...	...	...	...

Tab. 5.10 – Paramètres préférentiels suggérés à partir des tables de performances dédiées à l'apprentissage pour les paires `NPR-NCOM` : extrait du modèle de préférences complet proposé en annexe D.3.2

NPR-NCOM

Rang	**Critere**	**Poids** $[0,1]$	**Seuil pref.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	agrGen	0,28	-	-	-	1
2	agrNum	0,23	-	-	-	1
3	isDefinite	0,15	-	-	-	1
4	distExp	0,08	0,18	0,11	-	0,22
5	distTer	0,07	0,11	0,06	-	0,09
...	...	...	...	...	...	...

Afin de quantifier la pertinence de ces modèles de préférences suggérés à partir des observations sur les sous-tables de performances dédiées à l'apprentissage, nous les avons appliqués sur les sous-tables de performances dédiées à l'évaluation. Le tableau 5.12 illustre les résultats

TAB. 5.11 – Paramètres préférentiels suggérés à partir des tables de performances dédiées à l'apprentissage pour les paires NPR-PRON : extrait du modèle de préférences complet proposé en annexe D.3.2

NPR-PRON

Rang	Critere	Poids $[0,1]$	Seuil pref.	Seuil indiff.	Seuil veto	Limite
1	distSent	0,23	0,21	0,08	0,45	0,51
2	distExp	0,19	0,15	0,12	0,25	0,3
3	isClosestCandidate	0,15	-	-	-	1
4	isSubj	0,12	-	-	-	1
5	countOcc	0,06	1	0	-	1
...	...	...	...	...	...	...

obtenus à l'aide de ces modèles.

TAB. 5.12 – Résultats obtenus en utilisant les modèles de préférences suggérés par les méthodes statistiques

	NPR-NPR	NPR-NCOM	NPR-PRON
Précision	0,9	0,43	0,84
Rappel	0,92	0,52	0,44
F-mesure	0,91	0,47	0,58

À travers une analyse des valeurs proposées par les méthodes statistiques pour les différents paramètres décisionnels, nous avons demandé à l'expert de réviser ou de compléter ses modèles de préférences initiaux. Le troisième type de modèle de préférences ainsi obtenu, que nous avons qualifié de mixte, repose donc à la fois sur le résultat d'observations statistiques, mais également sur des intuitions émises par l'expert.

Cette mise en parallèle des modèles de préférences experts et observés a permis d'identifier à la fois les imprécisions des intuitions émises par l'expert, mais également les limites récurrentes des méthodes statistiques. En effet, nous avons constaté dans ce cadre d'expérimentation, mais également de manière générale, qu'un expert dispose souvent de connaissances précises sur la façon dont les critères de comparaison disponibles doivent être exploités. Nous avons cependant également observé qu'il était parfois délicat pour un expert de projeter avec précision ses intuitions sur un ensemble de paramètres numériques. Ainsi, les principales remises en cause des modèles de préférences experts concernaient la redéfinition des valeurs des paramètres décisionnels en s'appuyant sur ceux suggérés par les méthodes statistiques. Par rapport aux modèles de préférences suggérés, l'intervention de l'expert a permis d'identifier les phénomènes minoritaires non perçus par les méthodes statistiques. Les méthodes statistiques permettent d'identifier avec beaucoup de réussite des phénomènes majoritaires présents dans un corpus d'apprentissage, mais se retrouvent cependant inefficaces face à des phénomènes minoritaires pouvant pourtant être discriminants. Cette observation constitue une limite récurrente des approches statistiques, mettant ainsi en avant l'intérêt des méthodes mixtes ou hybrides. À titre d'exemple, on constate à l'aide de l'annexe D.3.2 que la propriété d'acronymie est estimée comme peu discriminante par les méthodes statistiques car elle n'apparaît qu'à très peu de reprises dans les sous-tables

de performances dédiées à l'apprentissage. Cependant, dans le cadre d'une paire d'expressions composée d'entités nommées, l'observation d'une propriété d'acronymie permet d'en déduire un lien de coréférence entre les expressions avec une assez grande certitude.

Ainsi ce travail de mise en parallèle d'un modèle de préférence observé et d'un modèle de préférences expert permet d'exploiter les généralités observées par les méthodes statistiques sur un ensemble représentatif d'exemples mais également les connaissances de l'expert permettant notamment de prendre en compte également des phénomènes minoritaires importants.

Généralement, les méthodes statistiques permettent d'établir des valeurs de paramètres décisionnels plus précis que les valeurs émises *a priori* par l'expert. Nous remarquons cependant que les préférences émises intuitivement par l'expert sont en majorité confirmées par les résultats des méthodes statistiques, ce qui nous permet d'attester de la pertinence des heuristiques sur lesquelles reposent ces méthodes.

Les tableaux 5.13, 5.14 et 5.15 illustrent les valeurs associées aux paramètres décisionnels des modèles de préférences mixtes sur les cinq critères jugés comme prépondérants à la fois statistiquement et manuellement par l'expert.

Tab. 5.13 – Paramètres préférentiels issus de l'alignement entre les valeurs suggérées par les heuristiques statistiques et les connaissances expertes : extrait du modèle de préférences complet proposé en annexe D.3.3 pour les paires `NPR-NPR`

NPR-NPR

Rang	**Critere**	**Poids** $[0,1]$	**Seuil pref.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	wordsInCommon	0,25	0,2	0,0	0,0005	0,21
2	subString	0,25	-	-	-	1
3	isSimil	0,2	-	-	-	1
4	strSimil	0,2	0,26	0,12	0,54	0,71
5	isAcronym	0,1	-	-	-	1
...	...	...	...	...	...	...

Tab. 5.14 – Paramètres préférentiels issus de l'alignement entre les valeurs suggérées par les heuristiques statistiques et les connaissances expertes : extrait du modèle de préférences complet proposé en annexe D.3.3 pour les paires `NPR-NCOM`

NPR-NCOM

Rang	**Critere**	**Poids** $[0,1]$	**Seuil pref.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	agrGen	0,19	-	-	-	1
2	agrNum	0,15	-	-	-	1
3	appo	0,09	-	-	-	1
7	isDefinite	0,09	-	-	-	1
4	attributive	0,08	-	-	-	1
5	isIndef	0,07	0,0	0,0	0,01	1
...	...	...	...	...	...	...

TAB. 5.15 – Paramètres préférentiels issus de l'alignement entre les valeurs suggérées par les heuristiques statistiques et les connaissances expertes : extrait du modèle de préférences complet proposé en annexe D.3.3 pour les paires NPR-PRON

NPR-PRON

Rang	Critere	Poids $[0,1]$	Seuil pref.	Seuil indiff.	Seuil veto	Limite
1	distSent	0,18	0,21	0,08	0,45	0,51
2	distExp	0,15	0,15	0,12	0,25	0,3
3	isClosestCandidate	0,12	-	-	-	1
4	isSubj	0,09	-	-	-	1
5	countOcc	0,05	1	0	-	1
...	...	...	...	...	...	...

Le tableau 5.16 montre que des résultats intéressants sont obtenus par l'application de la méthodologie de contrôle par tri associée aux modèles de préférences mixtes.

TAB. 5.16 – Résultats obtenus en utilisant les modèles de préférences mixtes

	NPR-NPR	NPR-NCOM	NPR-PRON
Précision	0,92	0,53	0,73
Rappel	0,95	0,47	0,53
F-mesure	0,93	0,50	0,61

La comparaison des résultats obtenus avec ces trois types de modèles de préférences différents (expert, suggéré et mixte) nous permet de valider l'importance d'une démarche hybride reposant à la fois sur des connaissances expertes et des observations statistiques. On constate en effet que les résultats obtenus à l'aide des modèles de préférences mixtes nous a permis d'atteindre de bons résultats en terme de précision et de rappel. Nous reviendrons sur l'explication de ces résultats dans la section dédiées à l'interprétation de ce cas d'expérimentation (section 5.2.3).

En soi, les résultats illustrés par les tableaux précédents ne sont pas réellement significatifs de l'efficacité de notre approche par surclassement sur le problème étudié. Nous avons donc exploité une méthode plus répandue pour ce genre de problématique, les arbres de décision [McCarthy and Lehnert, 1995], pour disposer d'une base de comparaison. Nous avons utilisé la méthode c4.5 proposée par l'API WEKA[19]) pour dans un premier temps construire trois arbres de décision à partir des sous-tables de performances dédiées à l'apprentissage, que nous avons dans un second temps évalués sur les sous-tables d'évaluation. Les résultats obtenus par cette méthode (tableau 5.17) sont étonnamment nettement moins bons que ceux obtenus à l'aide des méthodes d'AMCD par surclassement. Il faut tout de même souligner que ces résultats ne sont sans doute pas représentatifs de la qualité des méthodes basées sur des arbres de décision, dans la mesure où nous avons appliqué "naïvement" une méthode générique sur nos données, sans procéder au préalable à une étude approfondie de l'adéquation de cette méthode vis-à-vis de nos données. De meilleurs résultats pourraient sans doute être obtenus par l'intermédiaire de cette méthode en procédant à une optimisation des paramètres qu'elle exploite. De plus, nous avons souligné que les tables de performances, qu'elles soient dédiées à l'apprentissage ou à l'évaluation, contiennent une

[19] http://weka.sourceforge.net

faible proportion d'exemples positifs par rapport aux exemples négatifs. Cette forte dissymétrie rend délicat l'application d'une méthode d'apprentissage automatique et constitue sans doute la principale explication de l'infériorité des résultats obtenus à l'aide des arbres de décision.

TAB. 5.17 – Résultats obtenus en utilisant les arbres de décision

	NPR-NPR	NPR-NCOM	NPR-PRON
Précision	0,93	0,75	0,52
Rappel	0,9	0,29	0,43
F-mesure	0,91	0,42	0,47

Apport de la prise en compte d'un modèle de préférences lors de l'exploitation des critères

Lors de la première phase d'expérimentation, nous avons mis en avant les avantages et la pertinence de la démarche hybride que notre méthodologie offre à travers la prise en compte à la fois d'intuitions expertes et d'observations statistiques. Nous avons vu que cette démarche nous avait permis de construire un modèle de préférences mixte reflétant tout d'abord les phénomènes observés sur les sous-tables de performances dédiées à l'apprentissage, mais également les connaissances d'un expert utilisées pour compléter le modèle suggéré automatiquement.

Dans cette seconde phase d'expérimentation, toujours sur la problématique de la résolution des cas de coréférence, nous allons évaluer l'apport des approches par surclassement par rapport aux méthodes plus simples d'agrégation de critères. Nous allons ainsi montrer que la prise en compte d'informations précises sur les critères, matérialisées par les modèles de préférences, permet d'exploiter plus finement la connaissance apportée par chacun des critères concernés.

Ainsi, en s'appuyant sur un unique jeu de données constitué par les sous-tables de performances dédiées à l'évaluation, nous allons mesurer à l'aide de métriques identiques (précision, rappel et f-mesure) la performance de différentes stratégies d'agrégation. À travers l'énumération ci-dessous, nous verrons que nous avons dans un premier temps exploité uniquement les vecteurs de performances associés aux paires d'expressions candidates pour constituer les classes d'hypothèses valides et invalides. Puis, nous avons progressivement intégré des connaissances supplémentaires (sélection, poids et seuils) sur les critères pour arriver finalement à une méthode d'agrégation basée sur un modèle de préférences complet.

Méthode additive

Nous avons dans un premier temps normalisé l'ensemble des valeurs des tables de performances sur l'intervalle $[0, 1]$, afin de procéder à l'application d'une méthode additive simple. Cette méthode d'agrégation directe permet d'obtenir un critère unique synthétisant l'ensemble des performances associées à une hypothèse h_i. Cette valeur obtenue, notée G_{h_i}, est ensuite comparée à un seuil d'acceptabilité, noté S, validant ou non l'hypothèse en tant que coréférence :

$$G_{h_i} = \sum_{j=1}^{m} g_j h_i$$

$$h_i \in C_1 \text{ si } G_{h_i} \geq S$$

où C_1 correspond à la classe des hypothèses valides.

Pour obtenir ce seuil global d'acceptabilité, noté S, nous avons calculé cette performance synthétisée pour chaque exemple des tables d'apprentissages et nous avons déterminé le seuil permettant de mieux discriminer les exemples positifs des exemples négatifs.

Le seuil d'acceptabilité obtenu a ensuite été utilisé pour valider ou invalider les hypothèses des sous-tables d'évaluation par rapport à la valeur représentant l'agrégation de leurs vecteurs de performances. Nous rappelons qu'ici G_{h_i} correspond simplement à la somme des performances obtenues par les hypothèses sur l'ensemble des critères concernés. Les résultats obtenus à l'aide de cette méthode additive sont présentés dans le tableau 5.18.

TAB. 5.18 – Résultats obtenus en appliquant une méthode additive

	NPR-NPR	NPR-NCOM	NPR-PRON
Seuil d'acceptabilité	12,82	6,4	7,9
Variation de G_{h_i}	négatif $[1,2-14,8]$ positifs $[6,9-19,1]$	négatif $[1,1-9,2]$ positifs $[2-10,3]$	négatif $[1,2-14,8]$ positifs $[1,9-13,6]$
Précision	0,8	0,27	0,39
Rappel	0,92	0,91	0,92
F-mesure	0,86	0,42	0,54

.

Bien que les résultats obtenus par cette méthode naïve d'agrégation par addition directe des performances soient relativement corrects, ils reposent tout de même sur une transformation des données initiales pour les rendre commensurables. En effet, la normalisation de l'ensemble des données sur un intervalle commun repose soit sur la disponibilité d'une telle fonction de normalisation, soit sur la performance maximale envisageable sur chacun des critères.

Méthode additive pondérée

Nous avons ensuite complété la méthode précédente par l'intégration de poids sur chacun des critères. La distribution de poids utilisée est celle générée par l'application de l'algorithme RELIEF sur les tables d'apprentissage (voir annexe D.3.2). La méthode de détermination du seuil global d'acceptabilité est identique à celle présentée précédemment. Étant donné que $\sum_{j=1}^{m} w_i = 1$, et que $\forall j, 1 \leq j \leq m, g_j(h_i) \in [0,1]$, G_{h_i} varie dans l'intervalle $[0,1]$. Les résultats obtenus à l'aide de cette méthode sont présentés dans le tableau 5.19.

$$G_{h_i} = \sum_{j=1}^{m} w_i.g_j(h_i)$$

$$h_i \in C_1 \text{ si } G_{h_i} \geq S$$

où C_1 correspond à la classe des hypothèses valides.

TAB. 5.19 – Résultats obtenus en appliquant une méthode additive pondérée

	NPR-NPR	NPR-NCOM	NPR-PRON
Seuil d'acceptabilité	0,68	0,25	0,63
Précision	0,866	0,30	0,52
Rappel	0,92	0,80	0,57
F-mesure	0,90	0,43	0,54

.

Méthode par vote

À l'aide de cette troisième méthode, nous avons cherché à nous affranchir du biais introduit par la normalisation des valeurs sur un intervalle commun et par la recherche d'un vecteur de données commensurables. Nous avons alors utilisé une méthode par vote basée sur la comparaison des performances avec des limites d'acceptabilité établies pour chaque critère. La détermination de ces limites $L : \{l_1, l_2, ...l_m\}$ a également été effectuée sur les tables d'apprentissage, où nous avons simplement cherché pour chaque critère la valeur de son domaine de définition qui permettait de discriminer au mieux les exemples positifs des négatifs. La valeur synthétisant l'ensemble des comparaisons est égale à la somme des poids des critères sur lesquels les performances de l'hypothèse surpassent les limites d'acceptabilité. Généralement, les hypothèses évaluées selon cette méthode de votes égalitaires sont considérées comme valides si une majorité $\frac{\sum_{j=1}^{m} w_j}{2}$ des critères utilisés s'accorde sur cette assertion de validité. Cependant, en fonction du type des paires d'expressions évaluées, le nombre de critères pertinents peut varier et cette notion de majorité absolue n'a plus de sens. Nous avons donc utilisé les tables d'apprentissage pour déterminer un seuil, correspondant au poids des votes positifs qu'une hypothèse doit obtenir pour être valide. Comme le montre la formule ci-dessous, G_{h_i} représente le nombre critères de l'hypothèse h_i qui surpassent les limites d'acceptabilité associées aux différents critères. Étant donné que ces votes sont pondérés et que $\sum_{j=1}^{m} w_j = 1$, G_{h_i} varie dans l'intervalle $[0, 1]$. Les résultats obtenus à l'aide de cette méthode sont illustrés dans le tableau 5.20.

$$G_{h_i} = \sum_{j=1}^{m} w_j \ . \ P_j h_i$$

où,

$$\begin{cases} P_j h_i = 1 & \text{si } g_j(h_i) \geq l_j \\ P_j h_i = 0 & \text{si } g_j(h_i) < l_j \end{cases}$$

TAB. 5.20 – Résultats obtenus en appliquant une méthode par vote

	NPR-NPR	NPR-NCOM	NPR-PRON
Seuil d'acceptabilité	0,63	0,5	0,42
Précision	0,99	0,43	0,52
Rappel	0,71	0,52	0,65
F-mesure	0,83	0,47	0,58

Méthode par surclassement : modèle de préférences mixte

Par rapport à la méthode précédente, les approches par surclassement, telles que la méthode ELECTRE TRI employée pour ce problème précis, introduisent à la fois des informations plus précises sur les différents critères (seuils de préférence, indifférence et veto) et la prise en compte de l'importance de la minorité des critères qui ne valident pas l'hypothèse (mesure de discordance). Le tableau 5.21 reprend les résultats obtenus précédemment suite à l'application d'une procédure de tri par surclassement basée sur un modèle de préférences mixte. On constate que les informations préférentielles introduisant des notions d'imprécision et d'incomparabilité permettent d'atteindre des résultats globalement meilleurs qu'en utilisant les autres méthodes d'agrégation.

TAB. 5.21 – Résultats obtenus en utilisant le modèle de préférences mixte

	NPR-NPR	NPR-NCOM	NPR-PRON
Précision	0,92	0,53	0,72
Rappel	0,94	0,46	0,54
F-mesure	0,94	0,50	0,62

Cette deuxième phase d'expérimentation sur la problématique de la résolution des cas de coréférence nous a permis de mettre en évidence l'apport des modèles de préférences dans une stratégie de contrôle.

5.2.3 Interprétation des résultats : propriétés linguistiques des reprises anaphoriques

Vers une meilleure compréhension du phénomène étudié

Cette expérimentation nous a permis à la fois d'avoir un regard critique sur l'approche de contrôle par surclassement que nous avons proposée, mais également sur le phénomène de la coréférence d'un point de vue linguistique. Nous avons pu dans un premier temps attester de l'utilisabilité de notre méthodologie à travers son application sur un cas concret d'indétermination. La première phase d'expérimentation nous a notamment permis de constater qu'un expert disposait souvent de connaissances précises et importantes sur le phénomène étudié et que ces connaissances pouvaient être formulées par l'intermédiaire d'un modèle de préférences. Ces connaissances *a priori* sont principalement concentrées autour de l'importance des différents critères utilisés, mais également la manière dont ils doivent être utilisés.

Les différents modèles de préférences émis par l'expert et illustrés par l'annexe D.3.1 traduisent ces connaissances. On constate notamment que la volonté de l'expert de traiter séparément les paires d'expressions extraites en fonction de leur type s'explique à travers la formulation de ces modèles de préférences, où les critères jugés comme discriminants varient en fonction du type de paire concerné. En effet, la sélection des critères exploités et la distribution des poids associée diffèrent radicalement en fonction des types de paires. Pour le cas des expressions contenant des noms propres, les critères mis en avant par l'expert étaient principalement d'ordre "typographique" (similarité `strSimil`, sous-chaîne `substring`, etc.) ou syntaxique (fonction sujet `isSubject`, parallélisme en fonction sujet `tightParallel`, etc.)[20].

Pour les paires d'expressions de type NPR-NCOM, il apparaît moins évident de dégager un sous-ensemble de critères fortement discriminants comme peuvent l'être les distances typogra-

[20] Voir l'annexe D.2 pour une description des critères.

phiques pour les paires NPR-NPR. C'est pourquoi on peut observer un plus grand équilibre dans la détermination des différents critères utilisés. Ces critères sont de différents types, syntaxiques (fonction sujet de l'antécédent, des deux expressions dans des phrases adjacentes, etc), contextuels (distance) et même "typographiques". Bien que fortement informatives, les informations de conformité du genre et du nombre sont peu exploitées. À travers une étude des informations disponibles, notamment lors de la construction du corpus de référence, l'expert a pu remarquer que les marques permettant d'identifier le genre et le nombre des entités nommées étaient peu fréquentes ("Mr", "Mme", "le", etc.).

Quant au cas des paires NPR-PRON, les critères de distance ainsi que de marques syntaxiques (présence d'un article défini ou d'un démonstratif, fonction sujet, etc.) ont été privilégiés par l'expert. Comme pour le cas des paires NPR-NCOM, il est délicat de faire ressortir un sous-ensemble de critères fortement discriminants, ce qui se traduit notamment par une distribution relativement homogène de poids sur les différents critères utilisés.

Confronter les connaissances expertes et les observations statistiques

Ces choix émis par l'expert lors de l'identification des critères prépondérants sont dans la plupart des cas confirmés par les distributions de poids proposés par la méthode RELIEF. On remarque en effet que les hiérarchies de critères obtenues par observation des tables de performances dédiées à l'apprentissage privilégient les même critères que ceux identifiés par l'expert. Cette vision statistique de l'usage des critères à utiliser pour identifier les paires d'expressions coréférentes nous a également permis de voir l'importance d'un traitement particulier pour chaque type de paires. On constate en effet que les critères identifiés comme discriminants par RELIEF diffèrent en fonction du type de paires traité.

Par rapport aux modèles de préférences proposés par l'expert on constate principalement que les poids proposés par RELIEF paraissent moins approximatifs, notamment par le fait que ces poids sont représentatifs du corpus utilisé. Ces poids sont en effet représentatifs du corpus d'apprentissage et sont donc potentiellement plus pertinents. On remarque cependant que les critères binaires, majoritaires dans la famille de critères exploitées, sont rarement jugés comme très discriminants par RELIEF. Cette observation nous a permis de constater une limite à l'usage de l'algorithme RELIEF pour l'évaluation et la quantification de l'importance relative des critères binaires utilisés. En effet, cette méthode est basée sur une mesure de distance entre les performances associées aux hypothèses valides et non valides. Or, ces critères binaires représentent des propriétés, souvent syntaxiques, qui ne sont représentatives que de certaines configuration de liens de coréférence (comme l'acronymie, certains pronoms démonstratifs, le parallélisme des fonctions syntaxiques, etc.). Ainsi, en ne constituant que des informations discriminantes pour des cas particuliers, les algorithmes basés sur la fréquence ou plus précisément la régularité et l'homogénéité peinent à considérer comme importante ce genre d'information (variation maximale de 1 entre les performances d'une même classe pour ces critères). Ce constat constitue une limite récurrente des approches statistiques qui sont certes très efficaces pour identifier les régularités observables sur corpus, mais qui restent cependant souvent muettes pour l'identification de phénomènes minoritaires. C'est pourquoi des critères intéressants tels que l'acronymie pour les paires NPR-NPR, l'apposition pour les paires NPR-NCOM ou les marques de défini et de démonstratif pour les paires NPR-PRON, ne sont pas associés à des poids importants par la méthode RELIEF.

La remarque précédente explique également pourquoi les critères incompatibles avec l'exis-

tence d'un lien de coréférence, tels que la présence d'une marque indéfinie ou l'absence de déterminant dans une expression nominale, n'ont pas été pris en compte. L'expert, en exploitant ses connaissances sur le domaine, exploite donc avec plus de pertinence ces informations, en déterminant notamment un seuil veto qui permet d'invalider une paire d'expressions qui posséderait ce genre de propriétés.

On constate également qu'il est délicat d'attester la pertinence des heuristiques statistiques proposées dans la section 4.3.3 dans la mesure où la plupart des critères exploités pour ce cas de contrôle sont de nature binaire. En revanche, pour les critères numériques de distance entre expressions ou les mesures de distance typographique, l'observation des courbes de répartition des exemples positifs et négatifs a permis d'identifier des valeurs possibles pour les seuils de préférence et d'indifférence, qui semblent pertinents à en juger les bons résultats obtenus à l'aide des modèles de préférences automatiquement suggérés (tableaux 5.12 page 121).

Comme nous pouvions le présager, les meilleurs résultats ont été obtenues en croisant des connaissances expertes et des observations réalisées sur corpus. Les profils qualifiés de mixtes, reprennent principalement les valeurs des paramètres préférentiels proposées pour les critères numériques et sont complétés par les connaissances de l'expert sur des phénomènes plus minoritaires, tels que les cas particuliers de veto.

Quant à la deuxième phase d'expérimentation, elle nous a permis de mettre en avant l'apport des modèles de préférences par rapport à l'exploitation des différents critères disponibles. Dans une démarche progressive, nous avons vu que plus on intégrait de connaissances sur la façon dont les critères devaient être exploités, plus on obtenait un tri efficace des paires extraites.

Conclusion d'expérimentation

Pour conclure sur ce deuxième cas d'expérimentation de notre approche de contrôle par surclassement, nous pouvons avancer que cette application sur un problème concret d'indétermination a contribué à valider les aspects fonctionnels de notre méthodologie. Outre ce point que nous allons reprendre au cours de la description de l'expérimentation suivante, nous avons surtout constaté que la lisibilité des résultats obtenus (modèles de préférences suggérés et décisions de classification émises) nous permettait d'acquérir plus de connaissances sur le problème étudié. Ces connaissances concernent principalement l'identification des informations pertinentes à prendre en compte pour déterminer si un lien de coréférence existe ou non entre deux expressions. Mais les paramètres préférentiels suggérés tels que les seuils de préférence, d'indifférence et de veto nous offrent également une meilleure compréhension des caractéristiques des chaînes de coréférence. Par exemple, contrairement à des premières intuitions expertes, nous avons pu constater que le nombre d'occurrences d'une entité nommée ne constituait pas un critère utile pour déterminer si cette expression était un antécédent potentiel.

Bien que les méthodes statistiques, que nous avons proposées pour compléter notre approche lorsqu'un corpus de référence est disponible, nécessitent d'être évaluées plus en détail sur d'autres cas d'expérimentation comprenant plus de critères numériques, nous avons tout de même pu attester de leur utilité et la nécessité de croiser des connaissances expertes et des observations statistiques. En effet, nous avons vu que globalement la distribution des poids obtenus par `RELIEF` reprenait les préférences émises par un expert. Nous avons également constaté que les

autres paramètres préférentiels associés aux critères numériques étaient pertinents et surtout qu'ils permettaient à l'expert d'identifier des intervalles de valeurs pouvant être intéressants pour représenter des zones de préférence, d'indifférence ou de veto. L'application d'une méthode entièrement automatisée comme les arbres de décision nous a permis d'illustrer l'apport de la prise en compte de connaissances expertes. Nous avons en effet supposé que la faible proportion d'exemples positifs dans les tables de performances dédiées à l'apprentissage nuisait à la pertinence du modèle de classification inféré. L'utilisation de la méthode `ELECTRE TRI` que nous proposons, en tant qu'approche de tri par surclassement, s'appuie à la fois sur des observations statistiques effectuées sur les tables d'apprentissage et sur des connaissances expertes définies *a priori*, qui permettent notamment de prendre en compte des phénomènes de surface ou linguistiques difficilement observables sur un ensemble d'exemples peu volumineux.

5.3 Quelle transcription pour un SMS ? problématique de sélection

Lors des expérimentations précédentes, nous avons illustré l'application de notre approche de contrôle pour l'obtention d'un classement de rubriques d'indexation et d'un tri des couples d'expressions extraits d'un texte en fonction de l'existence ou non d'un lien de coréférence. Nous nous attaquons désormais à un processus initial d'analyse plus complexe, dans la mesure où il s'appuie sur l'application successive de plusieurs modules de traitement associés à des ressources linguistiques adaptées et volumineuses. Bien que nous ayons déjà introduit lors de la section 1.2.3 le processus de transcription de SMS[21], nous reviendrons dans cette section sur les étapes qui le composent. Nous verrons ensuite dans quelles conditions l'application des modules de traitement constitue des "points d'embarras" problématiques. Après avoir identifié l'origine des indéterminations et quantifié leur impact sur le processus de transcription, nous décrirons notre démarche de transformation des "points d'embarras" en "points de décision", puis la façon dont notre module de contrôle a été appliqué pour améliorer l'identification d'une meilleure transcription pour un SMS. Nous insisterons notamment sur l'apport de notre méthodologie et plus précisément sur la centralisation des informations décisionnelles et la généricité des outils de contrôle. Nous conclurons cette expérimentation en présentant les résultats encourageants obtenus et surtout les différentes perspectives d'amélioration évoquées au cours de cette expérimentation du contrôle du processus de transcription des SMS.

5.3.1 Le processus symbolique de transcription des SMS

Au cours de cette troisième expérimentation, nous allons nous intéresser au contrôle du processus de transcription de SMS réalisé par `TiLT` dans le cadre d'une application de vocalisation des SMS. Dans la mesure où ce processus d'analyse a déjà été décrit en détail dans la section 1.2.3 et par [Guimier De Neef *et al.*, 2007], nous n'esquisserons dans cette partie que les principaux aspects de ce traitement. La transcription de SMS effectuée par `TiLT` repose sur l'application successive de trois phases d'analyse :

1. segmentation du SMS :
   effectue un découpage et un typage (MOT, CHIFFRE, PONCTUATION, SMILEY) des segments identifés.

[21] Acronyme de Short Message Service.

2. analyse lexicale des segments :
   associe à chaque segment une ou plusieurs entrées lexicales et procède à d'éventuelles corrections adaptées au style SMS.
3. analyse syntaxique de surface :
   regroupe les unités lexicales en constituants. Cette étape d'analyse syntaxique permet de déterminer par effet de bord la succession d'unités lexicales la plus pertinente en fonction de règles syntaxiques.

Les trois étapes d'analyse du processus de transcription exploitent des ressources linguistiques adaptées à cette forme atypique de communication écrite [Bové, 2005]. L'étape de segmentation exploite un ensemble de règles de découpage, qui ont été enrichies pour la reconnaissance de formes particulières comme les `émoticônes`. L'analyse lexicale des segments construits est effectuée par consultation d'un lexique "classique" d'environ 100 000 entrées, auquel 2 000 abréviations spécifiques au langage SMS ont été ajoutées (comme : lol, msg, 2min, etc.). Quant aux règles d'identification des constituants syntaxiques de surface, elles n'ont été que très légèrement modifiées dans la mesure où la syntaxe des messages SMS respecte les constructions syntaxiques du français standard.

L'adaptation de ces ressources linguistiques s'est appuyée sur des observations effectuées sur des corpus représentatifs de SMS et de leurs transcriptions. Cette analyse des particularités du langage SMS a été réalisée à l'aide de deux corpus, l'un développé par le laboratoire `DELIC` [Hocq, 2006] (9 700 SMS et leurs transcriptions), l'autre par l'université catholique de Louvain [Fairon and Paumier, 2006] (30 000 SMS et leurs transcriptions). Des extraits de ces deux corpus sont proposés dans l'annexe E.1.

Avant de constater que le processus de transcription des SMS réalisé à l'aide de `TiLT` est également confronté au problème de la génération et de la propagation d'hypothèses erronées, il apparaît important de préciser la nature des hypothèses générées au cours de ce traitement et les interdépendances qu'elles entretiennent. Ce travail d'identification des indéterminations et de leurs effets sur le reste du traitement est indispensable, dans la mesure où il permet de saisir les particularités du contexte décisionnel (nombre d'hypothèses concurrentes, disponibilité de certains critères, etc.) et surtout de déterminer sur quels objets il est le plus pertinent d'appliquer une stratégie de contrôle. La figure 5.4 illustre graphiquement le processus de construction progressive des hypothèses linguistiques conduisant à un ensemble de transcriptions candidates, sans pour autant rentrer dans des détails liés à l'implémentation de `TiLT`.

On constate ainsi que l'application du module de segmentation sur le SMS d'entrée conduit à une liste de segments, où le comportement déterministe de ce module entraîne un découpage unique du message initial. Chacun de ces segments est ensuite transmis individuellement au module d'analyse lexicale. Issu d'un appariement direct avec des entrées du lexique ou de méthodes correctives, un ensemble d'hypothèses désignées sous le terme de terminaux est généré, où chaque terminal regroupe des informations lexicales, phonétiques, morpho-syntaxiques, etc. Les terminaux sont ensuite factorisés selon leur catégorie morpho-syntaxique (pronom, nom commun, verbe, etc.) pour construire ce que nous nommons des Groupes Syntaxiques de niveau 1 (`GS1`). L'application du module d'analyse syntaxique de surface exploite la distribution des `GS1` pour construire des constituants appelés Groupes Syntaxiques de niveau 2 (`GS2`). Cette étape de construction des `GS2` est basée sur l'application d'un ensemble de règles syntaxiques sur le treillis de `GS1` résultant du module d'analyse lexicale. Pour des raisons de complexité algorithmique, à la

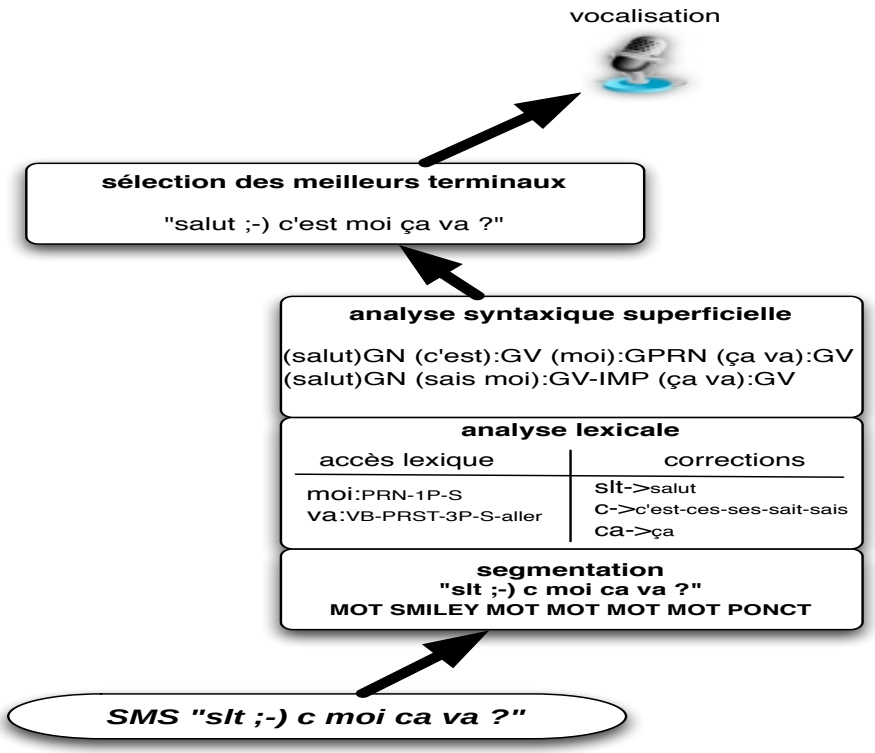

FIG. 5.4 – Processus de transcription des SMS

fois en terme de temps et d'espace, il n'est pas souhaitable de générer l'ensemble des distributions de GS2 valides selon les règles syntaxiques pour un treillis de GS1 donné. Ainsi en l'absence de critères de décision et de stratégie de contrôle pertinente, une heuristique forte inspirée de [Abney, 1991] entraîne la sélection du découpage possédant les plus longs constituants de gauche à droite. Comme nous allons le constater au cours du prochain paragraphe, cette heuristique n'est pas suffisante pour obtenir un comportement final déterministe. En effet, on observe, et particulièrement dans ce contexte d'analyse des SMS, que pour un découpage en constituants syntaxiques, il reste encore fréquemment plusieurs distributions de GS1 possibles. Imaginons par exemple le SMS suivant : "1 ptit rouge ca te tente" [22]. La stratégie de maximisation de la taille des GS2 de gauche à droite conduit à considérer "1 ptit rouge" en tant groupe nominal. Cependant, considérant ce premier découpage, il est toujours possible d'envisager deux analyses concurrentes, la première où "ptit - petit" est un nom commun (déterminant nom adjectif-attribut) et la seconde où "rouge" est un nom commun (déterminant adjectif-épithète nom). Il en est de même pour un constituant nominal tel que "les livres rouges", pour lequel "livres" est associé à deux terminaux possédant des traits de genre différents. Ces exemples restent artificiels, dans la mesure où les différentes hypothèses concurrentes conduiraient à une même vocalisation, mais de telles indéterminations ont lieu entre des terminaux de genre, de nombre ou de phonétisation différents entraînant ainsi des résultats complètement différents.

Pour traiter ces indéterminations, le processus actuel de traitement exploite un score associé à chaque terminal pour déterminer quelle transcription sera finalement proposée au service de vocalisation. Ces scores correspondent à des préférences numériques associées à des traits lexicaux

[22] Exemple construit purement artificiellement pour illustrer le phénomène de construction des GS2, non extrait des corpus DELIC et LOUVAIN.

ou morpho-syntaxiques, permettant ainsi de privilégier certains terminaux par l'intermédiaire des traits qui les qualifient. Parmi ces préférences émises par un expert linguiste, on remarque que ce critère basé sur des préférences *a priori* est utilisé notamment pour favoriser la prise en compte des locutions reconnues en affectant un score élevé aux terminaux correspondant à une locution. Les terminaux correspondant à des participes passés, visiblement peu fréquents dans les messages SMS, sont associés à un score faible. Le choix d'une transcription est alors basé sur la somme des scores des terminaux qui la composent. Lorsque cette information n'est pas discriminante, la première transcription générée est choisie.

On constate que le processus d'analyse montant jusqu'à l'identification des constituants syntaxiques permet une sélection de l'ensemble des terminaux construits par l'analyse lexicale. Cependant, malgré ce filtrage que l'on peut qualifier de "naturel" ou linguistique, le processus d'analyse se retrouve tout de même confronté à un "point d'embarras" qui a été dans un premier temps transformé en "point de décision" par la prise en compte d'un critère de score lexical. Comme nous allons le constater par la suite, cette stratégie a de nombreuses limites et la première est l'exploitation d'un seul critère pour résoudre un "point d'embarras" complexe.

5.3.2 Caractérisation et quantification des indéterminations

Origine des indéterminations

Sans revenir sur l'ensemble des difficultés liées au traitement de formes atypiques d'écriture comme le SMS [Véronis and Guimier De Neef, 2006], nous proposons à travers quelques exemples issus des corpus disponibles d'illustrer les origines de la génération des hypothèses concurrentes. Les quelques indéterminations que nous présentons apparaissent lors des différentes étapes du processus d'analyse, ce qui explique qu'au lieu d'être "naturellement" ou linguistiquement résolues, elles soient propagées ou même amplifiées.

Nous avons vu lors de la section 1.3 que l'analyse automatique d'énoncés écrits en français "standard" conduisait fréquemment à la génération d'hypothèses concurrentes et souvent erronées, ceci à cause de différents facteurs comme l'imprécision des ressources, l'ambiguïté inhérente des langues, les architectures trop rigides, etc. Le traitement de SMS, en tant que forme atypique d'écriture, introduit de nouveaux phénomènes à l'origine d'une augmentation du nombre d'hypothèses générées.

Pour illustrer les difficultés supplémentaires introduites par le traitement d'énoncés écrits en langage SMS, nous pouvons constater que l'analyse du segment très fréquent "c" conduit à de multiples interprétations :

- "voila c fini ca c b1 passé" ⇒ "voilà c'est fini ça s'est passé"
- "je c pa c ki" ⇒ "je sais pas c'est qui"
- "j esper k vou descendé c soir" ⇒ "j'espère que vous descendez ce soir"

On voit bien que contrairement à la plupart des contextes de traitement d'énoncés écrits en français standard, où l'interprétation lexicale des segments repose sur la consultation d'un lexique ou sur l'usage de méthodes correctives simples, le traitement des SMS nécessite d'avoir recours à des méthodes de correction, de déduction plus sophistiquées lors de l'interprétation des segments. Le segment "c" en est un bon exemple, mais on remarque également que l'identification des entités nommées devient également beaucoup plus complexe dans la mesure où chaque segment peut potentiellement être un nom propre : Faut-il interpréter le segment "mat"

comme un diminutif du prénom "Mathieu", la contraction du nom commun "mathématiques", de l'adjectif "mat", la conjugaison du verbe "mater", etc.

On remarque également que les SMS ne contiennent que très rarement des signes de ponctuation. On observe fréquemment des messages tels que : "alor t soulagé moi la j'aten 1h ca me soul je menui a mourir biz" ou "g un empechement previen sophie gros poutou à demain". Or, les signes de ponctuations sont des marques très importantes lors du découpage en constituants et l'écriture monolithique des SMS ne fait que rendre plus complexe la phase d'analyse syntaxique de surface.

Ces quelques exemples ne témoignent pas de l'ensemble des difficultés soulevées par cet exercice de transcription, mais contribuent à expliquer les chiffres présentés dans le paragraphe suivant visant à quantifier l'indéterminisme du processus étudié.

Quantification des indéterminations

À partir d'une analyse de 1000 SMS extraits aléatoirement des deux corpus, nous avons comptabilisé pour les différents types d'hypothèses le nombre moyen de candidats générés. Ainsi, on constate qu'en moyenne environ 29 segments sont construits par SMS. Pour chaque segment, le module d'analyse lexicale et ses méthodes correctives génèrent en moyenne 15 terminaux, qui sont ensuite factorisés sous 3 catégories morpho-syntaxiques concurrentes en moyenne. Pour chaque SMS, le module d'analyse syntaxique de surface construit en moyenne 3,8 constituants, dont chacun peut ensuite conduire à 1,7 analyses en moyenne.

Ces chiffres ne correspondent pas exactement au nombre des hypothèses concurrentes qui seront comparées par notre stratégie de contrôle. L'application du module d'analyse syntaxique a justement pour objectif de filtrer ces 15 terminaux concurrents par segment en fonction de leur validité syntaxique locale dans le constituant et globale dans la phrase. Nous verrons dans les parties suivantes que le nombre d'hypothèses concurrentes finalement comparées est bien moins élevé.

Évaluations du processus initial de transcription et identification des améliorations possibles

Le processus initial de transcription de SMS réalisé par `TiLT` a été évalué à deux reprises, une première fois sur le corpus du `DELIC` [Guimier De Neef *et al.*, 2007], puis plus récemment sur le corpus de `Louvain` [Guimier De Neef and Fessard, 2007]. Comme nous allons le constater, notre expérimentation s'appuie sur ces deux évaluations réalisées par les membres de l'équipe `Langues Naturelles` pour mettre en exergue la nécessité de compléter le processus initial de transcription par une stratégie de contrôle.

L'évaluation du processus de transcription de SMS réalisé par `TiLT` sur le corpus du `DELIC` avait comme principaux objectifs de disposer d'une première estimation de la pertinence des résultats générés, mais également de mieux comprendre les problèmes induits par ce traitement et ainsi d'identifier des perspectives d'amélioration.

Bénéficiant donc d'un corpus composé de 9 700 SMS et de leur transcription, cette première évaluation a été effectuée à l'aide de deux métriques : `Jaccard` et `BLEU`, pour mesurer la distance entre les transcriptions de référence et les transcriptions générées par `TiLT`. Ces deux mesures

évaluent la distance entre deux groupes de mots, `BLEU` exploite des `n-grammes` pour prendre en compte l'ordre, contrairement à `Jaccard` qui peut être qualifiée de mesure ensembliste. Ces deux mesures, qui seront reprises pour évaluer l'apport de notre stratégie de contrôle sur ce processus initial de transcription, sont calculées de la manière suivante :

$$\text{Coefficient de Jaccard} = \frac{|\text{intersection des mots}|}{|\text{union des mots}|}$$

$$\text{BLEU} = BP.exp\frac{1}{N}\sum_{n=1}^{N}\log(\frac{\text{nb. occurrences n-grammes}}{\text{nb. n-grammes}})$$

où :

- nb. occurrences n-grammes est le nombre de `n-grammes` communs avec au moins une référence[23],
- nb n-grammes est le nombre de `n-grammes` possibles dans la phrase $= taillephrase - N - 1$, N étant la taille maximale du n-gramme
- BP correspond à une pénalité infligée lorsque l'hypothèse est plus courte, en nombre de mots que la plus petite transcription de référence. $BP = min(1; exp(1 - \frac{\text{nb. mots référence}}{\text{nb. mots hypothèse}})$

L'application de ces mesures sur les hypothèses de transcription générées par `TiLT` a permis de constater que 25% des 9 700 SMS étaient parfaitement retranscrits et que globalement les résultas obtenus étaient satisfaisants (tableau 5.22).

Tab. 5.22 – Évaluation du processus initial de transcription de `TiLT` sur le corpus du `DELIC`

Jaccard	BLEU
0,749	0,712

Une analyse des 25 à 30% de transcriptions erronées proposées par `TiLT` a permis d'identifier les principales sources d'erreur. La première concerne les messages agglutinés, où aucune marque de segmentation n'est disponible : "Bonnefeteprofitesbiendevotredernierjourdevacance". Sans recours à des techniques particulières de segmentation, ces messages ne peuvent pas être analysés. Le module de segmentation étant le seul module de traitement de `TiLT` déterministe, aucune amélioration en terme de contrôle ne peut être envisagée sur ce point.

L'étude des erreurs d'interprétation a également soulevé d'intéressantes perspectives. Il a été constaté que des successions de mots fréquemment observées dans le corpus de transcription, telles que "rien de spécial" ou "comment fait-on ", n'étaient pas présentes dans les hypothèses finalement générées par `TiLT`. Ce constat ouvre donc des perspectives vers l'usage des corpus de transcriptions de référence pour effectuer un apprentissage et une identification des successions de formes récurrentes. Avant d'envisager la mise en place d'une stratégie de contrôle allant dans ce sens, il faut évidemment se poser la question de savoir si l'absence de ces successions fréquentes est due à des imprécisions des ressources linguistiques utilisées ou à une limite de la stratégie de sélection d'une meilleure transcription, stratégie relativement simpliste que nous avons présentée

[23] Plusieurs transcriptions de référence peuvent êre associées à un message SMS initial, voir [Guimier De Neef *et al.*, 2007] pour plus de détails.

précédemment (section 5.3.1). En effet, l'usage de stratégie de contrôle n'a de sens que si des hypothèses pertinentes existent parmi l'ensemble des hypothèses générées. Comme nous allons le constater, la deuxième évaluation effectuée sur le processus initial de transcription de `TiLT` nous permettra de répondre à cette interrogation.

Afin de vérifier la stabilité du comportement de `TiLT`, une seconde évaluation a donc été effectuée sur le corpus de `Louvain` [Fairon and Paumier, 2006]. Ce corpus est composé de 75 000 SMS, dont 30 000 ont été manuellement transcrits en français. Nous n'avons conservé que 10000 SMS pour l'évaluation du processus de transcription. Cette sélection nous permet de disposer d'un sous-corpus d'évaluation et d'un autre pour l'usage de méthodes statistiques d'apprentissage. Les résultats commentés dans l'article [Guimier De Neef and Fessard, 2007] montrent la stabilité de la fiabilité des transcriptions générées par `TiLT`.

Tab. 5.23 – Évaluation du processus initial de transcription de `TiLT` sur le corpus de `Louvain`

`Jaccard`	`BLEU`
0,745	0,736

L'analyse des résultats de cette seconde évaluation a permis d'identifier les mêmes sources d'erreur, à savoir la segmentation de SMS sans séparateur de mots et des erreurs liées à la stratégie de sélection d'une meilleure transcription.

L'article [Guimier De Neef and Fessard, 2007] décrivant cette évaluation ne présente pas l'ensemble des études effectuées sur le corpus de `Louvain`. En effet, dans le cadre d'un stage de `master 2`, Sébastien Fessard, étudiant de l'université de Provence, a effectué une étude très détaillée des erreurs commises par `TiLT` lors de la transcription des SMS du corpus de `Louvain`. Ces travaux avaient notamment comme principal objectif d'identifier l'origine des transcriptions erronées. Plus précisément, il s'agissait de distinguer les cas où les erreurs provenaient d'incomplétudes des ressources linguistiques (lexiques, grammaire de chunking, ou stratégies correctives) des erreurs liées à une mauvaise sélection des unités lexicales à utiliser pour la transcription. Cette distinction est cruciale pour l'application de notre approche de contrôle. Si les erreurs proviennent d'incomplétudes dans les ressources linguistiques utilisées, alors cela signifie que l'ajout d'une stratégie de contrôle n'apporterait rien. En effet, l'amélioration des résultats reposerait alors sur l'acquisition de ressources linguistiques (lexiques et règles de constructions syntaxiques) supplémentaires et plus adaptées, et non sur une remise en cause de la stratégie de sélection des meilleurs unités lexicales participant à la construction des transcriptions. Il est évident que si individuellement les transcriptions correctes des différents segments du SMS ne sont pas générées, jamais une transcription complète et correcte du SMS ne sera envisageable.

Si au contraire, cette étude montre que pour les SMS mal transcrits les unités lexicales correctes étaient générées par le module de traitement lexical, mais que le choix effectué par le module d'analyse syntaxique de surface s'est porté sur des unités erronées lors de la construction de la transcription, alors il devient envisageable d'améliorer les résultats à l'aide d'une stratégie de contrôle plus pertinente. Augmenter la précision générale des résultats de ce processus d'analyse reposerait alors sur une meilleure sélection des unités lexicales (terminaux) lors de la phase d'analyse syntaxique.

L'analyse des sorties complètes[24] de TiLT a montré que 7 942 SMS du corpus Louvain, soit environ 25%, étaient mal transcrits alors que les segments des SMS étaient couverts lexicalement, et 2080 SMS du corpus du DELIC, soit environ 21%, étaient également couverts lexicalement mais pour lesquels la transcription finale générée n'était pas complètement correcte. Ainsi, l'interprétation de cette première source d'erreur nous permet d'envisager une marge d'amélioration en modifiant la stratégie décisionnelle de sélection des terminaux à prendre en compte lors de la construction finale de la transcription.

Afin de classer les erreurs et d'identifier les étapes du traitement sur lesquelles il sera nécessaire de focaliser notre attention lors de la mise en place d'une stratégie de contrôle, une comparaison entre le découpage en constituants proposé sur les messages SMS et le découpage proposé sur leurs transcriptions est effectuée. Les transcriptions manuelles générées par l'université de Louvain pour les SMS du corpus constituent des énoncés courts en français "simple" et correct. En considérant que les résultats générés par le module d'analyse syntaxique de surface sur de tels énoncés sont fiables et globalement corrects, il devient alors possible de les comparer aux hypothèses syntaxiques générées par ce même module sur les SMS. Cette comparaison nous permettra de déterminer si notre démarche de contrôle doit se focaliser sur le contrôle des terminaux issus de l'application du module lexical ou plutôt sur la stratégie de découpage en constituants syntaxiques.

Cette analyse a conduit à une catégorisation des différentes erreurs observées en fonction du nombre de constituants construits, de la distribution des catégories de GS2 sur le découpage en constituants (groupe nominal, verbal, adjectival, prépositionnel, etc.), la valence des constituants (nombre de mots).

1. isomorphie complète ;
   Le découpage en constituants, la taille des constituants et la distribution des catégories syntaxiques sont identiques. L'erreur provient donc d'une erreur de sélection des terminaux respectant l'analyse syntaxique construite.
2. malgré des nombres de constituants syntaxiques identiques, les analyses générées sur le SMS initial et les transcriptions de référence possèdent des distributions de catégories syntaxiques différentes.
3. malgré des nombres de constituants syntaxiques identiques, les analyses générées sur le SMS initial et les transcriptions de référence possèdent des valences[25] de constituants différentes (nombre de mots).
4. malgré des nombres de constituants syntaxiques identiques, les analyses générées sur le SMS initial et les transcriptions de référence possèdent des distributions de catégories syntaxiques et des valences de constituants différentes.
5. les analyses générées sur le SMS initial et ses transcriptions de référence ne possèdent pas le même nombre de constituants.

La figure 5.5 illustre la répartition des erreurs d'analyse syntaxique de surface commises par TiLT sur les 7 942 SMS du corpus Louvain et sur les 2080 SMS du corpus du DELIC pourtant couverts lexicalement.

[24] Sorties en XML contenant les traces de toutes les hypothèses intermédiaires construites : terminaux, GS1, GS2 et AGSN (Analyse de Groupe de Second Niveau, i.e. distribution de catégorie morpho-syntaxiques).
[25] Correspond au nombre de mots du constituant.

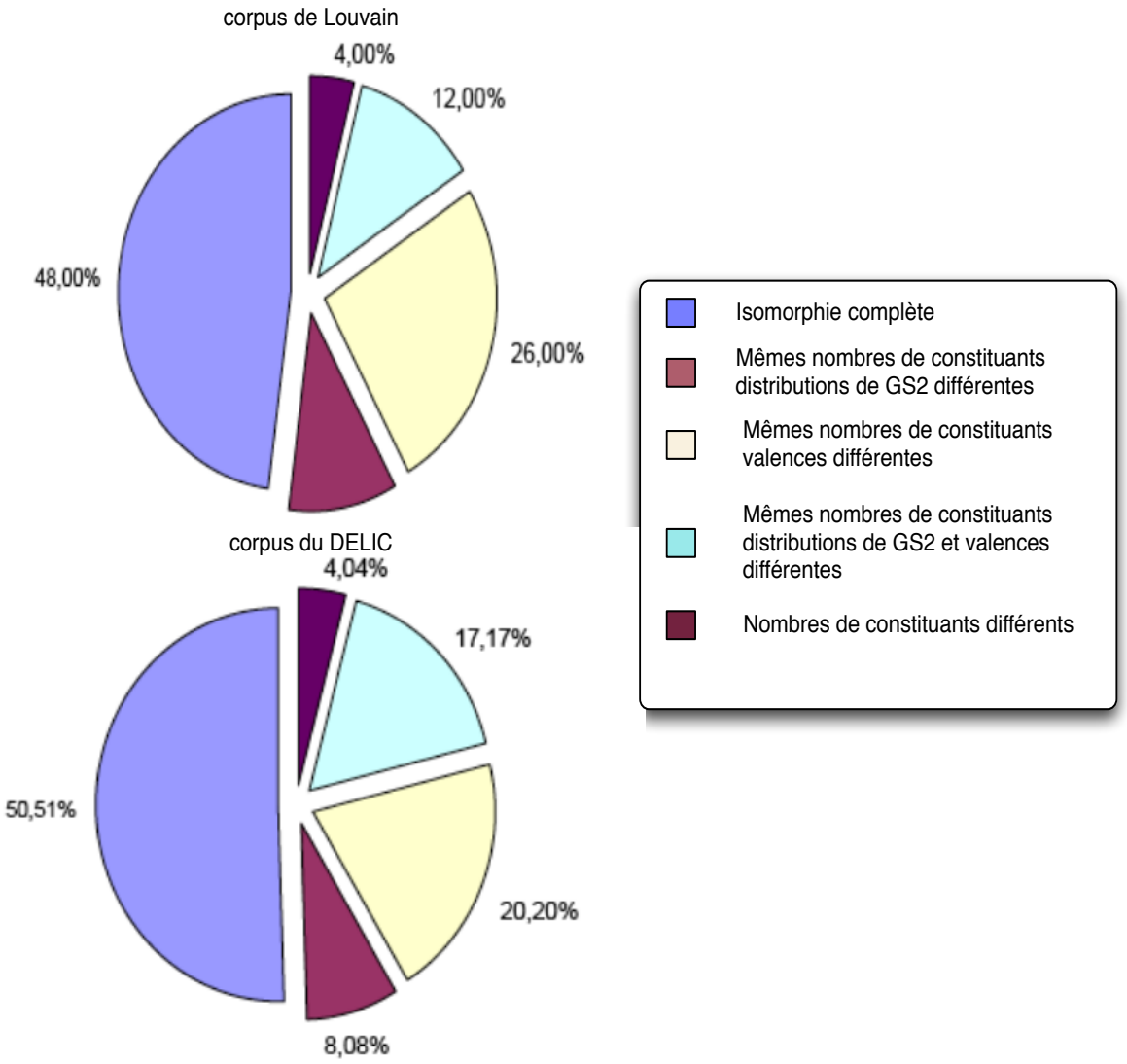

FIG. 5.5 – Répartition des erreurs syntaxiques identifiées sur les corpus de `LOUVAIN` et du `DELIC`

On constate que les répartitions d'erreurs syntaxiques commises sur les deux corpus sont relativement équivalentes. Parmi les SMS dont l'analyse n'a pas permis d'obtenir une transcription correcte, on observe que dans environ 50% des cas, les résultats de l'analyse syntaxique de surface obtenus sur une transcription de référence et ceux obtenus sur le SMS d'origine conduisent à des analyses syntaxiques de surface différentes. Cette observation montre que la stratégie de découpage basée uniquement sur la maximisation de la longueur des constituants de gauche à droite constitue une source d'erreurs potentielle. Cette stratégie de découpage est également à l'origine des erreurs caractérisées par des valences de constituants différentes.

Il semblerait donc que pour les 40% environ d'erreurs restants, la sélection des terminaux finalement exploités pour la construction d'une transcription soit fautive. Pour espérer améliorer le processus de transcription sur les 25% de SMS couverts lexicalement mais mal transcrits, il apparaît nécessaire de contrôler à la fois les processus d'analyse lexicale et syntaxique. Nous allons constater au cours de la sous-section suivante (5.3.3) que la mise en place d'une stratégie de contrôle du découpage en constituants syntaxiques conduit systématiquement au contrôle des différents terminaux concurrents générés lors de l'analyse lexicale.

Au cours de cette phase d'évaluation du processus initial de transcription de `TiLT`, une expérimentation intéressante a été menée. Suite à des modifications de la méthode de correction phonétique, une amélioration de la couverture lexicale a été observée sur le corpus du `DELIC`. En revanche, on constate que cette augmentation de 5% de la couverture lexicale a entraîné une régression (4%) de la précision finale des transcriptions générées. Cette observation confirme un constat que nous avons effectué lors du chapitre 1 section 1.3.2, concernant le rapport de symétrie entre couverture et précision. En effet, sans stratégie de contrôle efficace, le développement de ressources linguistiques plus larges se traduit par une génération d'hypothèses concurrentes supplémentaires. Pour cette problématique de transcription de SMS, l'usage d'un lexique plus volumineux ou de méthodes correctives plus ouvertes entraîne une sur-génération de terminaux rendant ainsi la phase d'analyse syntaxique plus complexe.

C'est pour cette raison que l'usage de méthodes correctives lors de la phase d'analyse lexicale est limité aux cas où la consultation du lexique ne permet pas de générer une analyse lexicale des segments. Cependant, les deux évaluations du processus initial de transcription ont montré que le traitement des homophones hétérographes était à l'origine d'erreurs fréquentes et que la seule façon de résoudre ces problèmes était l'usage systématique des méthodes correctives. En effet, au cours de la stratégie actuelle de traitement, le module d'analyse lexicale procède à des corrections et donc des suggestions de formes (terminaux) supplémentaires uniquement lorsque la forme du segment n'est pas reconnue dans le lexique. Cependant, dans un SMS tel que "ca se refus pas", le segment "refus" n'est associé qu'à une hypothèse de nom commun. La forme "refus" étant présente dans le lexique, les méthodes de correction ne sont pas appliquées et il sera impossible d'atteindre le terminal correct "refuse" en tant que verbe. Il apparaît donc d'autant plus primordial de compléter le processus d'analyse par des stratégies spécifiques de contrôle, afin d'augmenter à terme la couverture lexicale par l'application systématique des méthodes correctives.

5.3.3 Vers un contrôle statistique du processus d'analyse

Le processus de traitement conduisant à une analyse syntaxique de surface illustre parfaitement la problématique qui est à l'origine de nos travaux. Dès l'application des premiers traite-

ments, c'est-à-dire du module d'analyse lexicale, on constate l'apparition d'hypothèses concurrentes et erronées. Dans ce contexte de transcription de SMS, cette combinatoire apparaît comme légitime dans la mesure où l'analyse d'une forme initiale repose la plupart du temps sur des stratégies de suggestion ou de correction. Bien que l'application du module d'analyse syntaxique de surface ait pour but de déterminer les successions d'unités lexicales conformes d'un point de vue syntaxique, on constate que la propagation des indéterminations lexicales rend délicate et parfois inefficace l'étape de construction des constituants syntaxiques. En effet, la combinatoire présente dans le treillis des hypothèses lexicales ne fait qu'augmenter le nombre de découpages en constituants syntaxiques possibles. Comme la plupart des systèmes de traitement séquentiel, le module d'analyse syntaxique exploite une heuristique très forte (maximisation de la taille des constituants de gauche à droite) pour forcer le choix d'un découpage parmi tous ceux envisageables.

Le traitement des messages SMS en vue de leur transcription s'apparente donc à une succession de "points d'embarras", caractérisée par la propagation des indéterminations soulevées lors de l'analyse lexicale. La mise en place d'une stratégie de contrôle de ce processus nécessite tout d'abord de transformer ces "points d'embarras" en "points de décision" à travers l'intégration de critères de comparaison permettant d'évaluer la pertinence relative des différentes hypothèses concurrentes.

La mise en place d'une telle stratégie de contrôle nécessite une réflexion sur deux aspects. Le premier concerne l'identification des points de contrôle, c'est-à-dire l'identification des objets concurrents à comparer et le type de décision à prendre (tri, sélection, classement). La seconde, corrélée avec la première, concerne la nature des connaissances supplémentaires intégrées et intégrables dans le système afin de fournir les critères de comparaison nécessaires.

Identification des "points d'embarras" à transformer en "points de décision"

Lors d'une évaluation du processus initial de transcription, nous avons pu constater qu'environ 20 à 25% des SMS n'étaient pas correctement bien transcrits alors qu'ils étaient couverts lexicalement. Ainsi, malgré la présence d'unités lexicales correctes, c'est-à-dire correspondant à des éléments de la transcription de référence, une mauvaise succession d'hypothèses lexicales est favorisée par le module d'analyse syntaxique.

Bien qu'il apparaisse instinctivement légitime d'évaluer les hypothèses lexicales concurrentes afin d'écarter les unités erronées, on constate que l'intégration d'une phase de contrôle juste après l'application du module d'analyse lexicale n'est pas pertinente. En effet, la phase d'analyse syntaxique de surface a pour objectif d'identifier les unités lexicales permettant de construire un découpage en constituants valide vis-à-vis des contraintes de construction des groupes syntaxiques et des contraintes de compatibilité entre groupes. Il semble donc davantage pertinent d'effectuer un contrôle des unités lexicales ayant déjà été soumises à cette validation syntaxique.

Nous avons cependant également remarqué lors de l'évaluation du processus initial de transcription que la stratégie de découpage en constituants pouvait être à l'origine d'erreurs. La sélection finale des unités lexicales participant à la construction des transcriptions s'appuie sur ce découpage en constituants et sur cette phase de validation des contraintes syntaxiques. Il serait alors inutile pour de nombreux cas d'erreurs de procéder à une sélection des meilleures unités lexicales sur une analyse syntaxique incorrecte. Il apparaît donc intéressant d'étudier dans quelle mesure le processus de découpage en constituants et d'analyse de leurs propriétés pourrait bé-

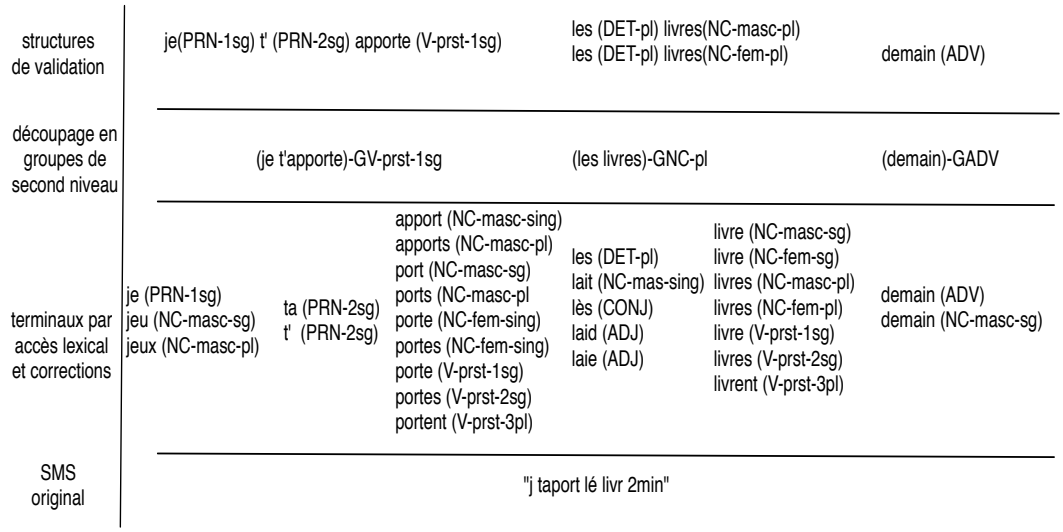

FIG. 5.6 – Enchaînement des hypothèses générées lors du processus de transcription

néficier de l'intégration d'informations supplémentaires et donc de la stratégie de contrôle. Pour cela, il convient de revenir sur le processus de construction des interprétations intermédiaires, sans pour autant rentrer dans des détails très complexes liés à l'implémentation particulière de `TiLT`.

En se rapportant à la figure 5.4 on constate que la première source d'indétermination résulte de l'application du module d'analyse lexicale. Les hypothèses lexicales générées, que nous nommons `Terminaux`, sont ensuite soumises au module d'analyse syntaxique de surface. À partir des contraintes syntaxiques inter-groupes permettant d'attester la validité d'une succession de constituants, objets linguistiques que nous nommons `GS2` pour groupe syntaxique de niveau 2, un découpage est établi en maximisant leur taille de gauche à droite. Les `GS2` ne constituent rien d'autre que des fenêtres de découpage de l'énoncé. Pour chaque fenêtre (i.e. `GS2`) établie lors du découpage, des objets nommés structures de validation, sont construits pour identifier les successions de terminaux du `GS2` valides selon les contraintes intra-groupes. Ces structures de validation correspondent à des restrictions syntaxiques sur l'ensemble des terminaux initiaux que nous avons évoquées précédemment. La figure 5.6 illustre sur un exemple très simple l'enchaînement de construction des hypothèses d'interprétation.

On constate donc que le processus d'analyse syntaxique est construit sur un ensemble d'objets linguistiques hautement corrélés et qu'il apparaît délicat de déterminer un point précis et indépendant de contrôle. En effet, pour espérer atteindre une sélection des meilleurs terminaux, il faut avoir au préalable identifié le meilleur découpage en `GS2` et pour chacun de ces `GS2`, il faut disposer d'une évaluation de la pertinence des différentes structures de validation possibles pour finalement effectuer la sélection des meilleurs terminaux restants.

Bien que ces différents objets linguistiques ne constituent que des vues différentes sur des terminaux (regroupement ou restriction), il apparaît nécessaire de procéder à un contrôle global et cohérent de ces différents types d'hypothèses générées lors du processus d'analyse de surface. Nous allons constater qu'en offrant une centralisation de tous les éléments décisionnels, notre module de contrôle permet une propagation et une réutilisation des critères de comparaison

disponibles pour l'évaluation de l'ensemble des hypothèses intermédiaires construites.

Critères de comparaison

Comme nous l'avons signalé précédemment, la mise en place d'une stratégie de contrôle repose également sur l'intégration ou la déclaration de critères de comparaison permettant de passer des "points d'embarras" aux "points de décision". Afin d'effectuer une sélection des terminaux finalement soumis à la construction de la transcription, le processus initial de traitement exploitait une information distinctive matérialisée par un score calculé lors de l'application du module d'analyse lexicale. Ce score, désigné sous le terme de `ScoreSyntaxique`[26], représente par l'intermédiaire d'une valeur numérique un jeu de préférences déterminé par un expert pour accorder plus ou moins d'importance à certains traits lexicaux, morpho-syntaxiques ou lexico-sémantiques portés par les terminaux. On constate que ce critère ne propose pas de discriminer finement la pertinence relative des différents terminaux et que cette stratégie de sélection des terminaux basée sur une maximisation de ce score est perfectible.

Nous avons dans un premier temps travaillé sur l'identification de sources de connaissances supplémentaires afin de compléter cette stratégie de sélection des meilleurs terminaux avec d'autres critères de comparaison.

Ce contexte d'expérimentation est notamment caractérisé par l'existence de deux corpus regroupant des SMS et leurs transcriptions de référence. Ces corpus constituent des sources de connaissances intéressantes sur lesquelles des méthodes statistiques peuvent être appliquées pour identifier des propriétés discriminantes récurrentes entre les hypothèses concurrentes.

Dans un autre cadre applicatif, celui de la campagne d'évaluation `GRACE`, une méthode statistique avait été développée afin de déterminer à partir d'un apprentissage de bi-grammes la succession de terminaux la plus fréquente parmi un treillis d'hypothèses. L'application de l'algorithme de Viterbi [Forney, 1973] sur un treillis d'unités lexicales permet de marquer le chemin le plus probable selon les fréquences d'apparition des couples de mots observées sur un corpus d'apprentissage.

Une partie du corpus de SMS collecté par l'université de Louvain (70%) a été exploitée pour construire une table de probabilités de bi-grammes de mots, permettant par exemple de déterminer la fréquence d'apparition du mot "te" précédé de "je" par rapport à toutes les occurrences de "te". Une fois le treillis de terminaux construit par le module d'analyse lexicale, cette table de probabilités est chargée pour déterminer la succession de terminaux la plus probable à l'aide de l'algorithme de Viterbi. Le résultat de cette évaluation statistique nous permet d'associer à chaque terminal construit un critère de type binaire correspondant à une marque d'appartenance ou non au chemin le plus probable. En complément de cette valeur binaire, nous ajoutons sur chaque terminal deux critères numériques déterminant la fréquence d'observation sur le corpus d'apprentissage de sa forme fléchie (conjuguée ou accordée) et de sa forme lemmatisée.

La présence de terminaux concurrents ne correspond ainsi plus à un "point d'embarras", mais bien à un "point de décision" caractérisé par la présence de quatre critères de comparaison représentant sous différents points de vue la pertinence de chaque terminal :

1. `scoreContrôleSyntaxique` : numérique ;

[26] Ce critère devrait être désigné par `ScoreLexicoSyntaxique` dans la mesure où le module d'analyse lexicale s'appuie sur des traits lexico-syntaxiques pour calculer ce score.

2. `appartientSolBigram` : binaire ;
3. `fréquenceFormeFléchie` : numérique ;
4. `fréquenceFormeLemmatisée` : numérique.

Cette première famille de critères, formée pour le contrôle des terminaux, est une illustration intéressante de la flexibilité de notre approche de contrôle. En effet, les méthodes par surclassement, en tant qu'approches multicritères permettant l'agrégation de performances hétérogènes de par leur nature et de par leur domaine de définition, nous offrent la possibilité de combiner des données numériques issues d'observations statistiques ou de préférences empiriques et des données binaires provenant de propriétés. Nous obtenons dans ce contexte précis une approche de contrôle parfaitement mixte (hybride), où des hypothèses linguistiques issues d'un traitement symbolique sont contrôlées par des connaissances principalement statistiques.

Intégration des phases de contrôle dans le processus d'analyse

La famille de critères que nous venons de décrire forme donc les éléments sur lesquels s'appuie la comparaison et la sélection des terminaux au sein d'une même structure de validation. Nous avons cependant vu que la phase finale de sélection des terminaux participant à la construction de la transcription n'était pas la seule source d'erreurs. La stratégie de découpage en constituants conduit également à des erreurs structurelles qui influent négativement sur la pertinence des résultats finaux. Les constituants, c'est-à-dire les `GS2`, ne sont en fait que des regroupements de terminaux validés par des contraintes inter-groupes (attestant la validité d'une succession de constituants). Nous avons alors cherché à exploiter les connaissances statistiques portées par les critères associés aux terminaux afin de compléter la stratégie de découpage en constituants.

Compte tenu du grand nombre de terminaux concurrents par `GS1` regroupés au sein d'un `GS2`, nous avons dans un premier temps fait remonter uniquement les connaissances relatives au meilleur chemin de bi-grammes construit à partir du treillis de terminaux. Il nous apparaissait en effet délicat d'agréger avec des méthodes naïves telles que la somme, la moyenne, le maximum ou le minimum, les autres critères de fréquence de forme ou de préférences lexico-syntaxiques tant les performances sur ces critères sont fortement disparates. Par exemple, sur des terminaux concurrents pour un segment initial donné, le nombre d'occurrences de la forme fléchie associée à chacun de ces terminaux peut varier de 0 à plus de 300.

Ainsi, lorsque le module d'analyse syntaxique construit un constituant de taille maximale, un critère nommé `GS2ScoreSolBiGram` lui est associé. Il correspond au pourcentage de ses terminaux qui appartiennent au meilleur chemin de bi-grammes. La validation d'un constituant repose désormais sur sa taille mais également sur le fait qu'il ne dégrade pas la meilleure solution de bi-grammes. Pour effectuer ce contrôle sur la base du critère `GS2ScoreSolBiGram`, nous soumettons chaque constituant proposé à une procédure de tri en exploitant l'opérateur par surclassement dédié à cette problématique. En considérant un profil d'acceptabilité décrivant les conditions de validité d'un constituant, nous pouvons soit conserver le constituant proposé, soit le remettre en cause et donc forcer le module d'analyse syntaxique à construire un nouveau constituant éventuellement de taille inférieure. Il est évident qu'une procédure de tri multicritère par surclassement ne semble pas d'un très grand intérêt pour déterminer la validité des constituants sur lesquels un seul critère est disponible. Cependant, cette mise en place d'une stratégie de contrôle externalisée et gérée par notre module de contrôle nous permettra de compléter facilement cette validation des constituants avec d'autres critères, notamment la prise en compte du meilleur chemin probabiliste de catégories morpho-syntaxiques dans le treillis de terminaux ou encore des

fréquences de constituants. Ces travaux sont en cours mais nous ne disposons pas à ce jour des résultats de cette nouvelle expérimentation.

Une fois le découpage en constituants établi, le module d'analyse syntaxique de surface construit les différentes structures de validation possibles sur les terminaux pour chaque `GS2` à partir des contraintes syntaxiques inter-groupes (conformité syntaxique entre les éléments du constituant : accord en genre, en nombre, etc.). La construction de ces structures de validation est une nouvelle source d'indétermination qui nécessite l'usage d'une stratégie de contrôle. Afin de déterminer la succession finale de terminaux qui constituera la transcription, nous avons intégré une étape de contrôle des structures de validation concurrentes dans le but de sélectionner la structure jugée comme la plus pertinente.

Lors de la section 4.2.4 nous avons présenté des fonctionnalités supplémentaires introduites dans notre système de contrôle. Nous avons notamment signalé le développement d'un opérateur d'agrégation donc l'objectif est d'associer à une hypothèse composite une vue agrégée des ensembles de critères associés à ses hypothèses de base. À l'aide de cet opérateur nous avons par exemple fait hériter chaque `GS1` des critères associés aux terminaux qu'il regroupe. Cette agrégation des critères des terminaux sur les `GS1` est possible à ce stade du processus d'analyse dans la mesure où une restriction des terminaux concurrents a déjà été effectuée lors de la construction des constituants (`GS2`).

Cet opérateur effectue donc une simple agrégation des critères des terminaux pour créer une famille de critères propre à chaque `GS1`. La configuration externalisée du module de contrôle `Beslissing` (voir section 4.1.4 et annexe B.2) détermine si les performances des critères du `GS1` doivent être calculées à partir d'une somme, d'une moyenne, du minimum ou du maximum des performances des critères des terminaux qui le constituent. Pour cette première expérimentation, nous avons choisi de calculer la valeur agrégée à l'aide d'une moyenne.

Une fois que la meilleure structure de validation a été identifiée pour chaque constituant, on procède à l'identification des meilleurs terminaux à l'aide également d'un opérateur de sélection par surclassement. Ces différents contrôles ne s'appliquent qu'en cas d'indétermination, nous avons observé que sur 1 000 SMS analysés (500 du corpus de `Louvain` et 500 du `DELIC`) nous avons en moyenne $1,7$ structures de validation concurrentes par `GS2` et $4,8$ terminaux concurrents en moyenne une fois la meilleure structure de validation sélectionnée. Ainsi, malgré un filtrage "naturel" des terminaux lors de la construction des constituants et donc de l'application de règles syntaxiques de surface, des indéterminations subsistent nécessitant une poursuite du contrôle.

Cette seconde expérimentation constitue un contexte intéressant d'évaluation de l'utilisabilité de notre module de contrôle décisionnel `Beslissing`. En effet, grâce à une modélisation générique, rigoureuse et fonctionnelle, nous avons pu déployer une stratégie de contrôle globale et cohérente sur un processus d'analyse complexe, ce qui auparavant n'était pas forcément envisageable, tout au moins pas aussi simplement. De plus, l'externalisation des paramètres décisionnels (voir section 4.1.4 et annexe B.2) nous laisse envisager des perspectives intéressantes quant à l'intégration de critères supplémentaires. On constate que la stratégie globale de contrôle mise en place offre la possibilité d'intégrer de nouveaux critères lors de la construction des unités lexicales et ensuite de les propager et de les exploiter pour contrôler la totalité du processus de transcription. La simplification de cette démarche de contrôle provient de la centralisation des éléments décision-

nels disponibles au cours d'un processus de traitement, mais également de l'externalisation du comportement du module décisionnel de contrôle dans un fichier de configuration.

La mise en place de cette stratégie de contrôle du processus de transcription des SMS a montré l'apport de la centralisation des données et outils décisionnels de contrôle, offrant ainsi à un expert la possibilité d'avoir une influence sur la qualité des interprétations générées tout au long du processus de transcription.

5.3.4 Application de la stratégie de contrôle, résultats et interprétations

Disposant d'une infrastructure globale permettant de contrôler les différents cas d'indétermination, il est désormais nécessaire de déterminer le modèle de préférences permettant une exploitation optimale des critères de comparaison disponibles. L'approche par surclassement dont nous nous sommes inspiré à travers l'usage des méthodes `ELECTRE III` et `ELECTRE TRI` s'appuie sur la formalisation de connaissances expertes pour former un modèle de préférences. La détermination des paramètres décisionnels pris en charge par notre méthodologie de contrôle par surclassement permet, comme nous l'avons remarqué lors de la précédente expérimentation, de déterminer l'importance relative des différents critères disponibles et d'introduire des notions d'imprécision et d'incomparabilité lors de la comparaison des performances de critères associées aux hypothèses concurrentes.

Bien que cette prise en compte de connaissances expertes *a priori* soit un avantage indéniable de notre approche, nous avons vu que sous réserve de disponibilité, un corpus de référence pouvait être exploité par des méthodes d'observation statistiques afin d'assister l'expert lors de la définition des différents paramètres décisionnels. Cette cohabitation de connaissances expertes et d'observations empiriques n'a cependant pas pu être mise en œuvre pour cette expérimentation autour du contrôle du processus de transcription des SMS. En effet, bien que nous disposions de corpus volumineux comprenant des SMS et leurs transcriptions, ces ressources ne constituent pas directement des corpus de référence pour nos différents points de contrôle. En effet, pour le problème de la résolution de la coréférence (section 5.2) le corpus de référence nous permettait directement de déterminer si une hypothèse (i.e. une paire d'expressions) et son ensemble de critères associé correspondait à un exemple positif (coréférence) ou négatif. Or, à partir des transcriptions qui ne sont qu'une succession de formes, il est impossible de dire si les objets linguistiques que nous contrôlons, c'est-à-dire les structures de validation et les terminaux, appartiennent à la référence. Il n'est donc pas possible d'établir directement les tables de performances (section 5.2.1) sur lesquelles sont appliqués l'algorithme `RELIEF` et les heuristiques statistiques de détermination des zones de préférence, d'indifférence et d'incomparabilité.

Nous avons donc dans un premier temps procédé à l'élaboration manuelle d'un modèle de préférences décrivant le comportement décisionnel des différents points de contrôle que nous avons intégré dans le processus d'analyse syntaxique. En concertation avec Christine Chardenon, responsable entre autres des aspects syntaxiques dans la chaîne de traitement `TiLT`, nous avons étudié quelques exemples de transcriptions erronées générées par `TiLT` et exploité nos connaissances et intuitions sur ce problème pour déterminer des valeurs initiales aux différents paramètres décisionnels pouvant être associés aux critères de comparaison. Nous avons notamment cherché à privilégier les critères statistiques comme l'appartenance à la solution de bi-grammes et les valeurs de fréquences.

Afin d'établir un modèle de préférences initial pertinent, nous avons procédé à une démarche

itérative au cours de laquelle nous avons fait varier les valeurs de ces paramètres puis évalué l'impact de cette variation sur les résultats. Nous avons construit un modèle initial avec des valeurs nulles pour les différents seuils, ce qui se traduit par une comparaison directe des performances obtenues par les hypothèses concurrentes sans considérer d'imprécision ni d'incomparabilité.

À partir de ce modèle initial, deux nouveaux modèles ont été construits, l'un intégrant une marge d'imprécision faible et l'autre une marge assez forte, ces marges étant obtenues en augmentant les seuils de préférence et d'indifférence. En nous appuyant sur une observation du domaine de variation des performances obtenues par les hypothèses concurrentes (terminaux et structures de validation), nous avons également proposé des seuils veto pour certains critères. Sur 1 000 SMS, nous avons comparé les résultats obtenus par ces différents modèles. Les valeurs des différents paramètres préférentiels qui nous ont permis d'obtenir les meilleurs résultats sur l'échantillon de SMS d'évaluation (1 000 SMS) ont été conservées pour construire le modèle de préférences utilisé pour le contrôle du processus de transcription.

Nous avons finalement cherché à valider notre intuition selon laquelle les critères statistiques apportaient des informations intéressantes pour discriminer les hypothèses concurrentes. Deux modèles de préférences supplémentaires ont été construits, le premier accordant une place très forte aux critères statistiques à travers une augmentation de leurs poids respectifs et une diminution du poids du critère de score lexico-syntaxique. Symétriquement, dans un second modèle de préférences, nous réduisons l'importance de ces critères statistiques au profit du critère de score lexico-syntaxique.

Suite à l'application de ces deux derniers modèles de préférences sur 1 000 SMS, nous avons constaté que les critères statistiques étaient bien les plus discriminants, mais que le score lexico-syntaxique apportait également un jugement intéressant sur la pertinence des hypothèses concurrentes à la fois au niveau des terminaux et des structures de validation. Par rapport aux résultats obtenus en exploitant notre modèle de préférences initial, ceux obtenus avec le modèle privilégiant très fortement les critères statistiques et inhibant l'impact du critère de score lexico-syntaxique sont nettement moins bons. Des résultats encore moins bons ont été obtenus à l'aide du modèle de préférences privilégiant le critère de score lexico-syntaxique au détriment des critères statistiques.

Cette simple expérience autour des modèles de préférences nous a permis de confirmer notre intuition selon laquelle les critères statistiques apportaient une discrimination intéressante des différentes hypothèses concurrentes. Nous avons également constaté que ces critères statistiques et le critère de score lexico-syntaxique étaient complémentaires et qu'ils devaient être combinés pour obtenir des résultats corrects.

Cette démarche progressive nous a conduit au modèle de préférences illustré par le tableau 5.24.

En ce qui concerne la procédure de tri des constituants permettant de déterminer si le meilleur chemin de bi-grammes est globalement respecté, nous avons établi une limite d'acceptabilité à $0,5$ avec tout de même une marge d'imprécision déterminée par les seuils de d'indifférence $(0,1)$ et de préférence $(0,3)$. Les constituants dont seulement 20% des terminaux appartiennent au meilleur chemin de bi-grammes ne sont pas acceptés (seuil veto $= 0,4$).

Cette stratégie de contrôle associée au modèle de préférences que nous venons de présenter a été évaluée sur les 10 000 SMS utilisés lors de l'évaluation du processus initial de transcription. En reprenant les mesures de `Jaccard` et `BLEU`, nous avons obtenu les résultats présentés dans le tableau 5.25.

TAB. 5.24 – Paramètres préférentiels définis par un expert pour la sélection des meilleurs structures de validation et des meilleurs terminaux

critère	poids	seuil pref.	seuil indif.	seuil veto
Sélection des structures de validation				
scoreContrôleSyntaxique	0,2	10	0	_
svaScoreSolBigram	0,4	0,2	0	_
fréquenceFormeFléchie	0,2	7	0	_
fréquenceFormeLemmatisée	0,2	10	0	_
Sélection des terminaux				
scoreContrôleSyntaxique	0,2	10	0	_
appartientSolBigram	0,4	0	0	_
fréquenceFormeFléchie	0,2	7	0	_
fréquenceFormeLemmatisée	0,2	10	0	_

TAB. 5.25 – Évaluation du processus contrôlé de transcription de TiLT sur le corpus de Louvain

Jaccard	BLEU
0,795	0,746

L'amélioration de la précision des résultats finaux observés sur ces mesures se matérialise par une diminution d'environ 5% du nombre de terminaux erronés finalement sélectionnés pour constituer les transcriptions. Le nombre de terminaux erronés dans les transcriptions proposées pour les 10 000 SMS du corpus d'évaluation passe de 31 248 à 29 759.

Afin de mieux comprendre cette progression qui *a priori* paraît faible, il semble intéressant de rapporter cette diminution du nombre d'erreurs à la marge de progression envisageable, en considérant la stratégie actuelle d'analyse et de contrôle. En effet, à travers une description des difficultés soulevées par l'analyse d'énoncés rédigés dans une forme d'écriture particulière comme peuvent l'être les SMS (voir section 5.3.2), nous avons avons vu que certains phénomènes tels que les agglutinations, les formes contractées très atypiques ou les mélanges de langues (français, anglais, franglais) ne pouvaient être gérés par le processus actuel d'analyse. De plus, nous avons vu que l'usage non systématique des méthodes correctives conduisait à des erreurs impossibles à rattraper par l'usage de stratégie de contrôle. Pour rappeler ce problème, nous pouvons prendre l'exemple de SMS suivant : "y son fou lé gas". La forme du segment "son" étant présente dans le lexique, le terminal correct correspondant à son interprétation en tant que verbe au présent de l'indicatif ne sera pas présent dans le treillis des terminaux.

Nous avons alors procédé à une estimation du nombre de terminaux erronés présents dans les résultats du processus initial de transcription pour lesquels nous pouvions espérer obtenir un résultat correct à l'aide d'une stratégie de contrôle. Cette marge de progression correspond au nombre de terminaux erronés dans les transcriptions générées pour lesquels un terminal concurrent correct était présent dans le treillis construit par le module d'analyse lexicale et les méthodes de correction. Ainsi, nous avons cherché à quantifier la marge réelle de progression possible pour disposer d'une appréciation plus juste de la qualité de notre stratégie de contrôle. Sur les 10 000 SMS de notre corpus d'évaluation, nous avons déterminé la proportion de mots des transcriptions pour lesquelles au moins un terminal avait une forme fléchie correcte. Nous avons ainsi

constaté que pour seulement 24% des formes erronées présentes dans les transcriptions générées par les processus initiaux d'analyse, un terminal correct était proposé par le module d'analyse lexicale. Cette marge de progression constitue une limite haute dans la mesure où ce comptage s'est effectué uniquement à partir d'une comparaison des formes. En effet, ce n'est pas parce qu'un terminal possède une forme fléchie similaire à celle attendue dans la transcription que l'on peut le qualifier de correct. Il peut en effet posséder une catégorie morpho-syntaxique ou des traits de genre et de nombre différents, ce qui le rendra certainement incompatible avec son voisinage syntaxique (contraintes inter-groupes lors de la construction des structures de validation) et l'écartera donc des terminaux finalement soumis à la procédure finale de sélection. Ces 24% constituent donc une marge de progression haute, qui n'est pas forcément accessible avec le système actuel d'analyse.

Ainsi, par rapport à cette marge de progression, l'apport de notre stratégie de contrôle devient beaucoup plus significatif. Nous avons en effet traité près de 20% des erreurs "à la portée d'une stratégie de contrôle".

On remarque que ces résultats encourageants ont été obtenus en utilisant un modèle de préférences ne disposant pas de seuils d'indifférence. Nous avons cherché à expliciter les raisons de cette particularité en analysant les résultats obtenus et également les critères associés aux structures de validation et aux terminaux comparés après la construction du découpage en constituants. Nous avons notamment constaté qu'une fois ce découpage en constituants obtenu, les indéterminations entre les terminaux restants n'étaient pas caractérisées par des formes fléchies différentes. En effet, les 4,8 terminaux concurrents comparés en moyenne une fois qu'une meilleure structure de validation a été sélectionnée ne possèdent que très rarement des catégories morpho-syntaxiques et des formes différentes. L'indéterminisme provient donc principalement de traits lexicaux tels que le genre ou le nombre, mais qui n'exercent pas d'influence sur la qualité des transcriptions. Pour appuyer cette observation, nous avons calculé, pour chaque sélection effectuée entre les terminaux concurrents restants, le nombre de cas où des distributions différentes de formes étaient présentes. Nous avons ainsi remarqué que dans seulement 3% des étapes de sélection des meilleurs terminaux, une ambiguïté de forme était encore présente. Or les quatre critères utilisés jusqu'ici exploitent des différences en terme de forme (fréquence et chemin de bi-grammes de formes) ou de catégorie morpho-syntaxique (score de préférence lexico-syntaxique) pour qualifier la pertinence des terminaux concurrents. Nous avons remarqué que les terminaux finalement comparés possédaient dans la plupart des cas des formes et des catégories syntaxiques similaires rendant ainsi les critères non discriminants. Ceci se matérialise par des performances de critères strictement équivalentes entre les terminaux concurrents. Cette observation explique donc pourquoi l'usage de seuils d'indifférence n'a aucun impact sur la pertinence des décisions (sélections) émises.

Ainsi, on constate que la réduction de 5% du nombre de formes erronées dans les transcriptions a été obtenue principalement en remettant en cause la procédure de découpage en constituants par un critère statistique de respect du meilleur chemin de bi-grammes. Ce constat nous permet d'orienter et de cibler nos travaux futurs concernant le contrôle de ce processus de transcription. Comme nous allons le constater par la suite, nous compléterons la procédure d'acceptation par le tri des constituants à l'aide de critères statistiques supplémentaires. Nous reconduirons également une procédure d'analyse des erreurs restantes similaire à celle présentée dans la section 5.3.2 et illustrée par la figure 5.5 pour confirmer que la part des erreurs de transcription due à un mauvais découpage en constituants a été réduite.

Perspectives et conclusion d'expérimentation

Le temps disponible pour mener cette expérimentation autour du contrôle du processus de transcription des SMS ne nous a pas permis d'appréhender tous les aspects soulevés par cette problématique. Les travaux menés ont tout de même conduit à la mise en place d'une stratégie complète de contrôle des différents cas d'indétermination, stratégie gérée entièrement par notre système de contrôle décisionnel.

Les résultats encourageants obtenus suite à l'utilisation de ces premiers critères statistiques ainsi que la mise en place de ce système de contrôle global nous laissent envisager des perspectives prometteuses, notamment à travers un usage plus prononcé des corpus disponibles. Nous avons précédemment signalé que les corpus ne pouvaient être utilisés directement pour construire les tables de performances nécessaires à l'application des méthodes statistiques d'aide à la définition des modèles de préférences. Nous travaillons actuellement sur la quantification du biais introduit par l'hypothèse qui consiste à considérer que chaque terminal dont la forme fléchie est présente dans la transcription de référence est correcte. En effet, cette hypothèse introduit un biais dans la mesure où différents terminaux concurrents peuvent partager une même forme fléchie, on parle alors d'homographes, sans pour autant être tous corrects d'un point de vue syntaxique (accord en genre et en nombre, terminaisons de verbes identiques pour des temps différents, etc.). En estimant ainsi la proportion d'homographes parmi les terminaux nous pourrons déterminer si cette hypothèse entraîne un biais significatif ou non. Si ce biais n'est pas important, nous pourrons alors construire des tables de performances regroupant les terminaux, leurs critères et leur évaluation (conforme ou non). Ces tables pourront ensuite être utilisées pour estimer statistiquement des valeurs à associer aux différents paramètres décisionnels.

De plus, suite à l'analyse des erreurs persistant malgré l'application de notre stratégie de contrôle, nous avons identifié des critères supplémentaires pouvant contribuer à l'identification finale des meilleurs terminaux. Nous avons notamment remarqué que les constructions syntaxiques utilisées lors de l'écriture de messages SMS variaient peu et que la plupart des transcriptions de SMS correspondait finalement à du texte "propre" dans un français relativement simple. Bien que la construction d'une analyse syntaxique de surface sur les messages SMS soit complexe en raison notamment de l'explosion combinatoire du nombre d'unités lexicales générées par le module d'analyse lexicale et les méthodes de correction, l'analyse syntaxique de surface de phrases courtes en français standard est efficace et les résultats souvent corrects. En réutilisant la méthodologie d'apprentissage de bi-grammes de mots et l'algorithme de Viterbi d'identification d'une meilleure succession de mots, nous travaillons actuellement à la construction d'une table de bi-grammes de catégories morpho-syntaxiques. Il sera donc ensuite possible de déterminer dans le treillis d'unités lexicales la meilleure succession de catégories morpho-syntaxiques et ainsi d'ajouter un critère binaire aux terminaux faisant partie de ce meilleur chemin. Afin d'avoir une estimation de l'apport de ce nouveau critère statistique, nous allons exploiter des tables de bi-grammes de catégories syntaxiques apprises sur un corpus de référence plus classique composé d'articles du journal `Le Monde`.

Nous pourrons ainsi compléter l'étape de validation des constituants, étape de contrôle qui apporte la principale amélioration de la précision du processus de transcription comme nous l'avons remarqué précédemment. Nous pourrons en effet valider un constituant par rapport à son respect des contraintes inter-groupes et sa taille maximale dans un premier temps, puis dans un second temps, vis-à-vis du meilleur chemin de bi-grammes de mots et celui de catégories syntaxiques. Nous compléterons cette validation par tri des constituants en ajoutant un critère dont

la performance correspond à la moyenne des fréquences de forme des terminaux qui les composent.

Nous avons également constaté que les bi-grammes n'étaient pas toujours suffisants pour reconnaître des patrons récurrents tels que "c'est moi" (trois mots "C" "EST" et "MOI"), "qu'est ce" ou encore "envoie moi le". Malgré la complexité engendrée par la prise en compte de tri-grammes, il serait intéressant de quantifier l'apport de ce nouveau critère pour l'identification des meilleurs terminaux en prenant en compte la pertinence statistique de leurs catégories morpho-syntaxiques au sein d'un voisinage syntaxique local.

De plus, nous avons signalé précédemment que l'usage d'une stratégie de contrôle permettait au maximum de réduire la proportion de formes erronées que de 24%. Pour atteindre un score global de 0,8 sur une mesure de `Jaccard`, il faut donc nécessairement augmenter de manière importante la couverture lexicale. L'obtention d'une couverture lexicale plus large nécessaire au traitement d'une plus grande partie du corpus nécessite l'usage systématique des méthodes de correction et de suggestion. Par exemple le SMS "ca vos rien", dont la transcription est "ça vaut rien" n'est pas traité correctement. En effet, étant donné que la forme "vos" est présente dans le lexique, aucune suggestion supplémentaire est proposée alors que le critère statistique de bi-grammes ainsi que les contraintes syntaxiques inter-groupes permettraient de choisir la forme correcte "vaut". La contre-partie à l'usage systématique de méthodes correctives est d'entraîner des sur-corrections se matérialisant par une régression de l'analyse de formes initiales correctes. Afin de contrôler ce phénomène, nous avons défini un nouveau critère représentant la confiance accordée à un terminal issu de méthodes correctives. Pour déterminer ce score de confiance, nous combinons la fréquence de la forme du segment initial et une distance typographique entre cette forme initiale et la forme du terminal. La performance obtenue sur ce critère serait forte pour les terminaux ayant une distance typographique élevée et une fréquence du segment initial faible, ce qui signifie que l'on cherche à corriger les segments dont la forme initiale est faiblement présente dans le corpus de transcription. Par exemple, la forme du segment "slt" est peu présente voire absente du corpus de transcriptions de référence, il est donc légitime d'accorder un score de confiance élevé aux terminaux qui la corrigent, en "salut" notamment. Symétriquement, les terminaux issus de méthodes correctives sur des formes de segments fortement présentes dans le corpus de transcription possèdent un score de confiance faible.

Il est évident que l'usage systématique des méthodes correctives entraîne une explosion combinatoire de la dimension des treillis de terminaux. Par intuition mais également de manière logique, cette explosion combinatoire se traduira par une chute importante de la précision des transcriptions générées par le processus initial non contrôlé de transcription. Nous pourrons alors une nouvelle fois montrer l'importance d'une stratégie de contrôle et sans doute observer une amélioration de la précision très significative à l'aide d'une validation multicritère de la pertinence symbolique et statistique des constituants et d'une sélection finale des terminaux concurrents.

On constate ainsi que notre système de contrôle centralisé, dont le comportement est externalisé dans un fichier de configuration, offre des perspectives d'expérimentations très intéressantes. En ajoutant simplement des critères sur les terminaux, nous pouvons à l'aide des opérateurs d'agrégation les propager sur d'autres hypothèses et les exploiter lors de différentes étapes de contrôle.

Malgré le peu de temps disponible pour mener cette expérimentation autour du contrôle du processus de transcription de SMS, nous avons obtenu des résultats très encourageants. L'interprétation des résultats obtenus nous a également permis de soulever des perspectives intéressantes

d'amélioration que nous allons développer à court terme.

Conclusion

Apports, perspectives et conclusion

1 Une intersection novatrice entre deux domaines de recherche

Comme l'a souligné Gérard Sabah dans [Sabah, 1989], le contrôle des cas d'indétermination, qu'il désigne clairement sous le terme de "point de décision", repose sur un processus rationnel d'exploitation de critères de jugement ou de comparaison conduisant à une prise de décision. Nous avons ensuite vu que l'exploitation de ces critères reposait sur la prise en compte de paramètres décisionnels supplémentaires, permettant notamment de déterminer l'importance relative de ces informations distinctives, mais également d'introduire des notions importantes telles que l'imprécision et l'incomparabilité.

La recherche d'une méthodologie conforme à ces constats nous a naturellement rapproché des méthodes issues de l'AMCD. De plus, nous avons vu au début du chapitre consacré à ces méthodes (chapitre 3) que notre contexte décisionnel remplissait les conditions nécessaires à l'application d'une méthode d'AMCD par surclassement et que les problématiques auxquelles ces méthodes répondent correspondent aux objectifs des stratégies de contrôle.

Cette intersection novatrice et expérimentale entre le TALN et l'AMCD constituait le cadre principal de recherche. Après avoir attesté de la faisabilité de la méthodologie proposée et après avoir évalué son utilisabilité et son efficacité sur différents cas concrets d'expérimentation, il nous faut désormais revenir sur les apports mutuels de nos travaux pour ces deux domaines de recherche.

1.1 Apport de notre approche pour la problématique étudiée

Lors du chapitre 2, nous avons décrit la propagation d'hypothèses concurrentes et potentiellement erronées au cours du processus d'analyse comme un phénomène problématique récurrent et difficile à résoudre uniquement à l'aide d'une adaptation ou d'une spécialisation des règles et des ressources linguistiques. Bien que ce problème ait été constaté sur de nombreuses stratégies d'analyse, sa résolution a toujours été considérée comme une correction *ad-hoc* ou une optimisation de la procédure initiale de traitement. En proposant à la fois une définition de la notion de contrôle et une terminologie orientée autour de notre vision décisionnelle du problème, notre approche a mis en évidence la nécessité de considérer la génération et la propagation d'hypothèses concurrentes comme un aspect induit par l'automatisation d'un processus cognitif complexe tel que l'analyse de données textuelles écrites en langue naturelle, mais également comme une caractéristique à part entière, certes négative, des systèmes de TALN. En rendant légitime l'usage de méthodes spécifiques de contrôle et en montrant leur apport sur des cas concrets d'indétermination, nous poursuivons l'idée de plus en présente dans les travaux récents en TALN que l'efficacité des procédures de traitement et la pertinence des interprétations qu'elles génèrent doivent être évaluées. En effet, depuis plusieurs décennies, les acteurs des développements en TALN se sont focalisés principalement sur les formalismes permettant de décrire les langues ou sur la construction de ressources linguistiques. L'objectif de la plupart de ces travaux était donc de faire progresser la robustesse des systèmes d'analyse et ainsi d'envisager de nombreux cadres applicatifs. Il semble désormais acquis que pour rendre l'ensemble de ces projets opérationnels et commercialisables, il devient indispensable de contrôler la pertinence des analyses générées. Nos travaux s'inscrivent donc dans cette démarche, mais en démontrant que de tels mécanismes de contrôle peuvent être intégrés de manière efficace et générique dans une architecture initiale de traitement, comme nous l'avons montré sur l'exemple de `TiLT`.

En nous appuyant sur plusieurs exemples, nous avons vu lors de la section 1.3 consacrée à la description des cas d'indétermination, que l'évaluation de la pertinence des différentes hypothèses concurrentes reposait sur la prise en compte de plusieurs critères de comparaison. Ces informations distinctives représentent différents points de vue ou niveaux de connaissances, dont chacun apporte une connaissance nécessaire mais rarement suffisante pour valider la pertinence d'une hypothèse. Bien que quelques travaux, en désambiguïsation du sens des mots notamment, aient montré la nécessité de cette prise en compte simultanée de plusieurs sources de connaissances, on constate que la plupart des stratégies de contrôle ne s'appuient que sur l'exploitation d'une seule source de jugement. En nous appuyant sur des méthodes issues de l'AMCD, nous proposons d'exploiter pleinement la complémentarité de ces différents critères et de gérer des phénomènes induits par ces procédures d'agrégation, tels que la compensation, la non-commensurabilité des informations et les contradictions.

De plus, l'usage de méthodes décisionnelles par surclassement permet de définir dans quelles conditions ces différents aspects doivent être traités. En effet, à travers la définition d'un modèle de préférences, un expert humain peut déterminer la manière dont les critères pris en compte doivent être interprétés et agrégés. Cette place prépondérante de l'expert humain lors de la construction d'une stratégie de contrôle est primordiale, car elle lui permet de formaliser des connaissances souvent pertinentes et suffisantes pour obtenir des résultats intéressants sur un problème précis.

Nous avons ensuite vu que l'utilisation de paramètres décisionnels numériques au sein de ces modèles de préférences nous permettait d'envisager des extensions statistiques résultant d'observations sur des corpus de référence, afin de faciliter la mise en place d'une stratégie décisionnelle de contrôle adéquate. Cette extension confère à notre approche une dimension hybride, qui comme nous l'avons vu lors de la section 5.2, permet d'exploiter de manière complémentaire les connaissances apportées automatiquement par les méthodes statistiques et celles définies *a priori* par l'expert. Ainsi, contrairement aux approches purement statistiques, notre approche s'appuie principalement sur la prise en compte de connaissances expertes.

Cette démarche que l'on peut qualifier de mixte ou hybride est intéressante à de nombreux égards. Elle permet tout d'abord de faciliter le travail de l'expert lors de la mise en place d'une stratégie de contrôle en lui suggérant des valeurs possibles. Elle conduit également fréquemment à des distributions de poids plus pertinentes, dans la mesure où elle reflète le contenu d'un corpus de référence représentatif. Mais cette extension statistique permet surtout de confirmer ou d'infirmer des intuitions formulées par un expert à partir de ses propres connaissances. Beaucoup de stratégies de contrôle s'appuient sur des heuristiques, mais sans qu'une validation de ces jugements empiriques soit effectuée. En comparant principalement l'importance accordée à un critère par un expert avec la distribution de poids obtenue automatiquement sur un corpus d'apprentissage, il est possible de déterminer la crédibilité à accorder aux critères manipulés et aux préférences formulées par l'expert.

De plus, cette remise en cause des connaissances définies et intégrées *a priori* par un expert est également rendue possible par la transparence et l'interprétabilité de la procédure décisionnelle réalisée par les méthodes `ELECTRE III` et `ELECTRE TRI` et des résultats qu'elle génère. En effet, il est aisé de déterminer par une simple observation des vecteurs de performances pourquoi une hypothèse en surclasse une autre. C'est également sur ce point souvent négligé que notre approche se démarque des méthodes plus fréquemment utilisées.

L'interprétation de la procédure décisionnelle, des modèles de préférences qu'elle exploite et des résultats qu'elle génère, permet notamment d'identifier des phénomènes distinctifs récurrents qui apportent des connaissances linguistiques importantes et intéressantes sur le problème étudié. Nous avons notamment pu constater lors de l'expérimentation sur la résolution de la coréférence, que les critères de distance n'étaient pas pertinents pour les paires d'expressions de type `NPR-NCOM`. De même, nous avons pu constater que les cas d'erreurs de tri pour les paires `NPR-NPR` s'expliquaient par l'absence de similarité typographique entre les deux expressions et que les critères disponibles actuellement ne permettraient pas de résoudre ces cas particuliers, reposant notamment sur des connaissances ontologiques ou universelles.

Bien que nos travaux puissent s'appliquer à de nombreux processus de TALN, particulièrement ceux suivant une architecture modulaire et séquentielle, l'étude, nos hypothèses et les développements sous-jacents se sont appuyés sur un exemple précis de chaîne : `TiLT`. Comme nous avons pu le constater lors du chapitre 4, nous nous sommes efforcés de construire une méthodologie de contrôle opérationnelle et de rendre cette proposition utilisable dans la version opérationnelle du logiciel `TiLT`. Il est donc évident que le principal bénéficiaire de ces travaux est l'équipe `Langues Naturelles` de `France Télécom division R&D`, qui dispose désormais d'un module opérationnel de contrôle pour son système de traitement linguistique. Dans un contexte de co-tutelle de ce projet entre un laboratoire universitaire et un laboratoire industriel, cet effort nous est apparu comme indispensable. Sans doute au détriment de recherches plus théoriques consacrées aux aspects linguistiques portés par les critères de comparaison utilisés, nous avons mis l'accent sur les aspects logiciels et informatiques de la mise en place d'une stratégie de contrôle. Ainsi, face aux différents cas d'indétermination observables lors de l'application d'une stratégie d'analyse à l'aide de la chaîne `TiLT`, les développeurs et linguistes de l'équipe `Langues Naturelles` disposent désormais d'une méthodologie fonctionnelle de contrôle. Ce module de contrôle que nous avons dénommé `Beslissing` permet de simplifier l'intégration et la manipulation de critères de comparaison ainsi que l'intégration de phases de contrôle au cours du processus d'analyse. L'application d'une stratégie de contrôle permet de disposer d'une évaluation de la pertinence relative des hypothèses évaluées, mais également de recommandations de décisions pouvant être exploitées pour augmenter la précision finale des résultats générés. De plus, la centralisation des informations décisionnelles et la traçabilité des procédures de contrôle permet d'obtenir une meilleure compréhension du domaine étudié. Cet aspect a notamment été mis en évidence lors du cas d'expérimentation sur la coréférence.

1.2 Apport de l'AMCD pour le TALN

Pour compléter la description proposée précédemment sur l'apport de notre approche vis-à-vis de la problématique étudiée, nous pouvons revenir plus précisément sur l'intérêt des méthodes d'AMCD pour le contrôle des cas d'indétermination et plus généralement pour le TALN.

Nous avons vu dans un premier temps que notre choix pour l'AMCD en tant que domaine d'inspiration pour la mise en place d'une méthodologie de contrôle était motivé par la volonté de disposer d'une approche spécialisée dans la résolution de problèmes décisionnels. Nous avons en effet posé que le contrôle pouvait se matérialiser par trois problématiques différentes, la sélection, le tri ou le classement et que la résolution de chacune d'elles reposait sur un processus décisionnel de comparaison soit des hypothèses entre elles (classement ou sélection), soit des hypothèses concurrentes avec des limites d'acceptabilité (tri). En nous appuyant sur une notion commune

de surclassement, les méthodes d'AMCD par surclassement proposent justement des processus d'élaboration de recommandation pour ces trois différentes problématiques. Ainsi, en s'appuyant sur cette notion générique de comparaison que constitue les relations de surclassement, nous avons pu concevoir et implémenter une méthodologie commune de contrôle applicable sur ces différentes problématiques.

Mais le principal apport pour notre problématique de l'AMCD et plus particulièrement de ces approches par surclassement, réside dans la prise en compte de connaissances et d'intuitions pouvant être formulées *a priori* par un expert du domaine. En effet, beaucoup de stratégies de contrôle sont basées sur l'hypothèse que tous les aspects discriminants nécessaires à la mise en place d'une procédure décisionnelle de contrôle pouvaient être observés statistiquement sur un corpus représentatif. En rappelant tout d'abord que la disponibilité de tels corpus constitue en soit une hypothèse forte, on constate également que ces approches ne laissent aucune place aux experts pour introduire leurs connaissances sur la manière dont l'opération de contrôle doit être établie. En s'inspirant de travaux théoriques et pragmatiques sur les processus décisionnels, les méthodes par surclassement en AMCD considèrent l'expert comme l'élément central de l'élaboration d'une recommandation de décision. En effet, les modèles de préférences regroupent l'ensemble des paramètres décisionnels pouvant influencer la procédure décisionnelle. Ainsi, en associant des valeurs à ces différents paramètres décisionnels, l'expert peut définir une stratégie de contrôle en accord avec ses connaissances et intuitions sur le cas d'indétermination concerné : choix des critères, importance relative, méthodes de comparaison (seuils), etc. En s'appuyant donc sur la définition des modèles de préférences proposés par les approches par surclassement et plus particulièrement par les méthodes `ELECTRE III` et `ELECTRE TRI`, nous profitons de leur maturité et de leur héritage des différents travaux portant sur la modélisation des préférences et la théorie de la décision. Nous pouvons ainsi directement focaliser notre attention et surtout celle des experts sur les aspects décisionnels prépondérant tels que les poids, seuils de préférence, d'indifférence et de veto et les limites d'acceptabilité.

L'AMCD apporte donc au TALN le cadre méthodologique permettant de tester et d'évaluer notre hypothèse selon laquelle la résolution des cas d'indétermination pouvait être formalisée comme un problème décisionnel basé sur l'exploitation de critères hétérogènes, potentiellement contradictoires et pas forcément commensurables. Les approches par surclassement nous permettent quant à elles de conserver cette propriété de disposer de méthodes de contrôle configurables, comme peut l'être `TiLT` pour la mise en place de stratégies d'analyse. C'est justement cette dernière propriété qui rend l'approche adaptable aux particularités des différents cas d'indétermination.

1.3 Apport du TALN pour l'AMCD

À première vue, l'apport de nos travaux et plus généralement du TALN pour le domaine de l'AMCD semble limité. En effet, nous n'avons fait qu'utiliser des méthodes relativement anciennes sans contribuer à leur amélioration, leur développement ou leur remise en cause. Cependant, comme nous l'avons signalé à plusieurs reprises au cours de ce document et comme nous allons le rappeler dans la section suivante, les cas d'application des méthodes décisionnelles de contrôle sont très nombreux en TALN. L'application de chaque module de traitement linguistique conduit potentiellement à l'apparition d'un cas d'indétermination. De plus, chacun de ces cas d'indétermination peut être traité ou complété par des critères de natures très variées. En proposant de réunir le TALN et l'AMCD autour de la problématique de contrôle, nous offrons

donc un grand nombre de cas d'application et de validation permettant de confirmer l'intérêt des méthodes issues des travaux en AMCD.

Plus particulièrement, les approches par surclassement se sont appuyées sur leur capacité de prise en compte des connaissances d'un expert lors de la mise en place d'une stratégie décisionnelle pour se différencier d'autres méthodes d'agrégation plus efficaces d'un point de vue algorithmique. La définition d'un modèle de préférences pertinent pour un contexte d'analyse et son éventuelle adaptation aux particularités d'autres contextes d'analyse constituent des exercices intéressants permettant de montrer l'utilisabilité de cette approche par surclassement et ses facultés à rendre compte des connaissances et intuitions pouvant être formulées par un expert sur le problème décisionnel étudié.

De plus, à travers l'usage de méthodes statistiques pour suggérer des valeurs possibles aux différents paramètres décisionnels, nous proposons une extension novatrice et intéressante aux méthodes d'AMCD par surclassement. Nous proposons en effet de confronter les préférences expertes définies *a priori* avec des observations statistiques effectuées sur un ensemble de décisions déjà émises sur le problème étudié. Bien que nous n'ayons à ce jour qu'effleuré l'ensemble des possibilités d'exploitation des corpus de référence pour la mise en place ou l'évaluation d'un modèle de préférences, nous avons tout de même démontré la pertinence et l'efficacité d'une telle approche hybride. Nous verrons également dans la section suivante que cet aspect sera au cœur de nos prochains travaux, notamment dans le cadre de la poursuite de la mise en place d'une stratégie de contrôle pour le processus de transcription des SMS.

2 Perspectives

Les travaux que nous avons menés n'avaient pas pour ambition de résoudre un problème aussi complexe que celui de la génération et la propagation d'hypothèses concurrentes et erronées. Nous avons cherché à poser un cadre descriptif et à la fois méthodologique pour ensuite l'intégrer dans un système existant de TALN. Bien que d'intéressants aspects aient été soulignés au cours de notre étude, de nombreuses pistes restent à explorer. Comme nous allons le constater, ces perspectives montrent que des améliorations méthodologiques peuvent être envisagées tout en conservant nos travaux théoriques et logiciels initiaux, à savoir considérer le traitement des cas d'indétermination comme des problèmes décisionnels et conserver le déploiement d'une méthodologie de contrôle en tant qu'élément à part entière d'une chaîne de traitement.

Bien que la rédaction de cette section soit frustrante dans la mesure où elle signifie que l'on s'oblige à figer les travaux effectués jusqu'ici, et à ne considérer ces pistes de réflexion intéressantes que comme des perspectives, dont l'étude reste encore conditionnelle, elle montre tout de même qu'une brèche a été ouverte et que nous ne nous trouvons pas dans une impasse.

2.1 Étendre le contrôle aux autres "points d'embarras"

Lors du chapitre 5, nous avons illustré le fonctionnement de notre module de contrôle décisionnel sur trois cas concrets d'indétermination. Ces expérimentations nous ont certes permis de valider l'intérêt de nos travaux, mais elles ont également soulevé certaines faiblesses, comme la difficulté d'évaluer la pertinence des valeurs suggérés par les heuristiques statistiques pour les seuils d'indifférence, de préférence et de veto. C'est pourquoi nous allons principalement focaliser nos prochains travaux sur ces aspects. Nous avons vu que le contrôle du processus de

transcription des SMS nous laissait envisager un tel couplage d'informations expertes et statistiques. Comme nous l'avons expliqué lors de la section 5.3.3, les corpus composés de SMS et de leurs transcriptions de référence ne peuvent être utilisés directement pour constituer des tables de performances comme cela a été le cas pour le problème de la résolution des liens de coréférence. À court terme, nous allons travailler sur l'exploitation de ces vastes corpus pour à la fois suggérer des modèles de préférences et également exploiter de nouveaux critères statistiques tels que les bigrammes de catégories morpho-syntaxiques. Mais l'exploitation de ce corpus nous permettra surtout d'avoir une évaluation du modèle de préférences défini *a priori* en croisant ces préférences avec des observations statistiques (usage de l'algorithme `RELIEF` et des heuristiques d'identification de zones de préférence, d'indifférence et de veto). Comme nous l'avons souligné dans la section 5.3.3, ces perspectives sont envisageables sous condition qu'un alignement puisse être fait entre les hypothèses concurrentes générées par `TiLT` et les transcriptions de référence.

Nous avons présenté lors du chapitre précédent trois cas différents d'indétermination sur lesquels il apparaissait nécessaire d'ajouter une stratégie de contrôle pour améliorer la pertinence et surtout la précision des résultats générés. Notre choix pour ces trois cas particuliers s'explique par le fait qu'ils constituaient déjà des "points de décision" et non des "points d'embarras". En effet, des informations distinctives pouvant jouer le rôle de critères de comparaison étaient déjà disponibles ou facilement récupérables grâce à notre module de contrôle. Cependant, la mise à disposition de notre module de contrôle pour les linguistes et informaticiens travaillant sur le développement de `TiLT` a suscité de nombreux besoins de contrôle et ceci à des niveaux différents de l'analyse. Parmi ces autres cas possibles d'application de notre méthodologie de contrôle, nous pouvons citer le processus d'analyse sémantique qui exploite les arbres de dépendances générés par le module d'analyse syntaxique pour construire un ensemble de graphes sémantiques candidats.

De plus, ces graphes sémantiques peuvent être par la suite exploités pour générer des traductions possibles de l'énoncé initial dans une autre langue. Ainsi, si l'on considère que l'analyse syntaxique génère un ensemble d'hypothèses concurrentes, que chacune d'ellee est interprétée par le module d'analyse sémantique pour générer un ensemble de graphes sémantiques concurrents et que chacun de ces graphes peut conduire à un ensemble de traductions possibles, on obtient une explosion combinatoire du nombre de branches d'analyse envisagé au cours du processus de traitement. Il serait donc très intéressant d'évaluer l'apport de l'intégration de phases de contrôle en différents points sur la précision finale dans les résultats proposés par un processus de traduction automatique, mais également de voir si le coût induit par l'intégration d'une phase de contrôle est inférieur au gain de temps obtenu par la suppression de certaines branches d'analyse. Ceci revient à vérifier si l'intégration d'étapes de contrôle intermédiaires nous permet d'arriver plus rapidement à une hypothèse finale correcte. Nous avons ainsi commencé à travailler sur le contrôle des arbres de dépendances concurrents générés par le module d'analyse syntaxique. Au cours de cette nouvelle expérimentation, nous allons chercher à identifier un sous-ensemble d'arbres de dépendances jugés comme les plus pertinents (problématique de sélection) et les propager ensuite au module d'analyse sémantique. En plus des critères associés aux terminaux qui évidemment composent également les arbres de dépendances, nous disposons de critères supplémentaires propres à l'évaluation de relations syntaxiques. Nous calculons en effet pour chaque arbre construit le nombre de cadres de sous-catégorisation qu'il remplit et symétriquement le nombre de cadres de sous-catégorisation qu'il ne remplit pas. Cette stratégie de contrôle est comparable à celle employée par l'analyseur `SYNTEX` [Bourigault, 2007].

Ces différentes perspectives d'application n'ont pas pu être développées pour le moment,

dans la mesure où elles nécessitent une réflexion préalable sur les critères de comparaison à intégrer ou les informations à considérer comme des critères de comparaison, pour notamment passer d'un "point d'embarras" à un "point de décision". Nous avons cependant déjà identifié des informations distinctives pouvant être traduites en critères, comme le test de la connexité des graphes sémantiques, l'héritage des critères associés aux analyses syntaxiques sous-jacentes, etc. Pour l'évaluation de la pertinence des traductions finales générées, différents travaux se sont déjà appuyés sur des sources de connaissances supplémentaires, telles que peut l'être INTERNET pour déterminer une meilleure traduction parmi toutes celles générées [Cao and Li, 2002].

Nos travaux se sont matérialisés par le développement d'un module de contrôle fonctionnel et intégré dans la version commerciale de `TiLT`, ainsi qu'une documentation technique d'usage, nous pouvons donc espérer que ces perspectives soient par la suite envisagées, dans la mesure où l'ensemble de la méthodologie est à la disposition de l'équipe `Langues Naturelles`.

2.2 Compléter l'approche et poursuivre son évaluation

L'application du module de contrôle pour l'identification des paires d'expressions coréférentes nous a permis de mettre en évidence l'intérêt d'une approche hybride et de l'exploitation des méthodes statistiques de suggestion de paramètres préférentiels que nous proposons. Nous avons cependant constaté que la grande proportion de critères binaires, type de critères sur lequel ces méthodes ne sont pas exploitables, ne nous permettait pas de déterminer complètement la pertinence des heuristiques proposées. De plus, en complément d'une expérimentation supplémentaire autour de la mise en place d'une stratégie de contrôle mixte, il serait intéressant de procéder à une analyse systématique de la sensibilité du modèle de préférences vis-à-vis des résultats obtenus. En faisant varier sur un certain intervalle les valeurs associées aux paramètres préférentiels et en évaluant l'impact de ces variations sur les résultats générés, il serait possible d'évaluer la stabilité d'un modèle et donc d'accorder plus de crédibilité à une stratégie de contrôle. Ainsi, la principale évolution à apporter à la méthodologie actuelle concerne une utilisation plus précise des corpus de référence disponibles, afin de disposer de modèles de préférences et d'une évaluation des intuitions expertes plus fiables.

De plus, nous avons commencé à mettre en place des scripts permettant de retracer à travers les logs générés les différentes étapes de l'application d'une stratégie de contrôle :

- instanciation des critères ;
- définition des structures de comparaisons ;
- construction des relations de surclassement ;
- interprétation des structures de préférences ;
- exploitation des recommandations de décision et propagation des hypothèses.

Il semblerait instructif de disposer d'une visualisation graphique simple de l'ensemble de ces étapes et ainsi pouvoir retrouver facilement les différentes hypothèses ayant contribué à la construction d'une interprétation finale, les critères qui leurs étaient associés, les résultats des comparaisons, etc. En complément des scripts disponibles actuellement, ces fonctionnalités de visualisation graphique seraient très utiles lors de la mise en place d'une stratégie finale de contrôle en s'appuyant sur l'observation des décisions émises par des stratégies temporaires de contrôle.

2.3 Extensions et pistes de recherches envisageables

Nous n'avons évidemment pas la prétention de considérer nos travaux comme une solution au problème complexe et récurrent de l'indéterminisme des chaînes de TALN. Nous avons cependant contribué à définir et cerner ce problème en considérant sa résolution comme un traitement à part entière qui doit être intégré dans le processus même d'analyse. Les travaux théoriques et méthodologiques que nous avons réalisés ont confirmé que l'analyse et plus précisément les aspects liés à l'évaluation des hypothèses concurrentes pouvaient être vus comme un processus décisionnel. En constatant l'adéquation entre les propriétés de notre problématique du contrôle des indéterminations et le paradigme décisionnel, nous avons cherché à exploiter des méthodes spécifiquement dédiées à la prise de décision ou plus précisément à la génération de recommandations de décisions. Les approches par surclassement sur lesquelles nos choix méthodologiques se sont portés ont montré de nombreux avantages, tels que la prise en compte de préférences expertes et l'interprétabilité des résultats générés. Nous avons cependant également relevé une limite majeure à l'application de ces méthodes dans notre contexte informatique de TALN. En effet, contrairement à la méthode `ELECTRE TRI` dont la complexité est linéaire en fonction du nombre d'hypothèses à trier, du nombre de classes considérées et du nombre de critères pris en compte, la méthode `ELECTRE III` que nous avons utilisée pour les problématiques de classement et de sélection a une complexité quadratique ($\frac{2(n-1)}{2}$ couples possibles dans un ensemble de n hypothèses). Cette forte complexité s'explique par la nécessité de comparer toutes les hypothèses concurrentes deux à deux. Cette obligation de comparer l'ensemble des hypothèses entre elles provient de la non transitivité de la relation binaire d'indifférence. En effet, l'introduction de seuils d'indifférence et de préférence permettant de tenir compte de la possible imprécision des performances de critères supprime la propriété de transitivité des relations de préférence et d'indifférence et donc de surclassement. Si par exemple on considère qu'une hypothèse h_1 est indifféremment préférée à une hypothèse h_2 ($h_1 I h_2$) et que cette hypothèse h_2 est indifféremment préférée à une hypothèse h_3 ($h_2 I h_3$), ceci ne nous permet pas d'établir que $h_1 I h_3$ puisque les différences de performances peuvent dépasser les seuils d'indifférence entre ces deux hypothèses sans pour autant les dépasser entre h_1 et h_2. La prise en compte d'un modèle de préférences rapproche les approches par surclassement de la réalité des contextes décisionnels, mais rend également de manière symétrique leur application plus complexe. Ainsi, en contre-partie d'une gestion de connaissances précise sur le fonctionnement du processus décisionnel, il apparaît obligatoire d'accepter une complexification de la procédure d'agrégation.

Sans remettre en cause notre vision décisionnelle du contrôle des processus de TALN, des compromis doivent être étudiés pour restreindre la prise en compte de paramètres décisionnels en faveur d'une réduction de la complexité des procédures de classement et de sélection. Une des pistes envisagées est d'étudier des méthodes décisionnelles alternatives plus récentes telles que les ensembles approximatifs [Greco *et al.*, 2001] qui ont notamment proposé de répondre aux mêmes problématiques que l'AMCD tout en réduisant la complexité des procédures de comparaison. De même, nous nous étions intéressés dans un premier temps à un opérateur d'agrégation défini par R. Yager [Yager, 1993], qui proposait une agrégation de multiples critères avec la prise en compte de la notion de compensation et donc de compromis pour des problématiques de classement. Étant donné que nous étions à la recherche d'une base méthodologique pouvant être appliquée pour les différentes problématiques envisageables, nous n'avions pas poursuivi cette piste pour nous intéresser plus précisément aux approches par surclassement. Il serait donc intéressant en complément à la méthodes `ELECTRE TRI` d'étudier l'usage de cet opérateur pour des problèmes de classement et éventuellement de sélection.

De même, la communauté travaillant sur la modélisation des préférences est très active et bien que les travaux dans ce domaine aient fortement influencé le développement des méthodes `ELECTRE`, une attention plus particulière pourrait être portée sur la modélisation des préférences, permettant notamment de représenter les dépendances entre les critères [Boutilier *et al.*, 2004][Dubois *et al.*, 2004] et d'arriver à des procédures d'agrégation beaucoup plus perfectionnées.

À court terme, nous allons enrichir les fonctionnalités du module de contrôle en proposant l'application d'opérateurs d'agrégation plus simples, notamment basés sur la comparaison de vecteurs de performances à l'aide d'un ordre lexicographique, pour traiter les cas où de nombreuses hypothèses concurrentes doivent être classées sans que la nécessité d'intégrer des paramètres préférentiels se fasse ressentir.

3 Conclusion

À travers ce document, nous avons cherché à décrire la démarche que nous avons menée pour traiter ou tout au moins étudier le problème particulier de l'indéterminisme des chaînes de traitement des langues. Au cours de ce projet, nous nous sommes efforcé de rendre cette démarche la plus scientifique possible, afin de garantir à la fois son intérêt et sa validité ainsi que sa potentielle réutilisation. En effet, comme en témoigne la structure de ce document, notre étude s'est dans un premier temps portée sur la description des cas d'indétermination et la définition d'un cadre théorique et terminologique adapté s'appuyant sur le paradigme décisionnel. Bien que nous ayons proposé une méthodologie dont nous avons montré l'intérêt sur trois cas concrets d'expérimentation, les travaux théoriques et la terminologie associée peuvent être réutilisés pour appliquer d'autres méthodes comme nous l'avons suggéré dans la section précédente 5.2.

Pour conclure cet écrit, nous proposons de revenir sur les différentes étapes qui ont composé cette démarche en nous intéressant notamment :

- à la description de la problématique de la génération et de la propagation d'hypothèses concurrentes, au cours des processus de TALN, hypothèses qui dans la plupart des cas correspondent à des erreurs d'interprétation ;
- à la formalisation du traitement des indéterminations en tant que problèmes décisionnels basés sur l'exploitation de plusieurs critères ;
- à l'appropriation des méthodes d'AMCD par surclassement pour constituer la méthodologie implémentée et intégrée dans `TiLT` ;
- puis à l'évaluation des hypothèses émises et à l'interprétation des résultats obtenus.

3.1 Légitimité des stratégies de contrôle

Nos travaux tirent leur légitimité du constat de l'incapacité des systèmes de TALN à contrôler le processus de génération d'interprétations linguistiques lors des différentes étapes de traitement. En effet, les recherches en TALN se sont longtemps focalisées sur la formalisation des connaissances linguistiques disponibles sur différentes langues, puis sur leur utilisation au sein d'une architecture logicielle de traitement. Les premières architectures de systèmes informatiques de TALN se sont inspirées de modèles issus des travaux en sciences cognitives, prônant un parallélisme et une interdépendance des ressources et des traitements linguistiques. Face à la complexité engendrée par cette transposition des modèles cognitifs dans des structures informatiques, la plupart des systèmes de TALN développés et maintenus jusqu'à ce jour a favorisé les aspects pragmatiques et fonctionnels au détriment des théories psycholinguistiques. Comme nous l'avons présenté lors du chapitre 1.1, les systèmes informatisés de TALN suivent principalement une architecture séquentielle et modulaire permettant une implémentation, un développement et une maintenance plus simples.

Comme nous l'avons illustré sur le cas de `TiLT`, la chaîne de traitement développée par `France Télécom`, les processus de TALN s'appuient sur l'application successive de modules de traitement dédiés à un niveau d'analyse précis, permettant au final d'atteindre le niveau d'interprétation souhaité. Ces systèmes informatiques de traitement se sont confrontés à un phénomène artificiel récurrent et fortement problématique, celui de la génération et de la propagation d'hypothèses concurrentes pouvant être erronées. En effet, bien qu'il soit légitime de considérer plusieurs interprétations pour un énoncé en entrée (voir section 1.3), l'application dans certains contextes applicatifs de systèmes de TALN comme `TiLT` conduit à un ensemble d'hypothèses concurrentes

dont une grande proportion est erronée. Longtemps, la résolution de ce problème s'est portée sur le développement de ressources linguistiques plus restrictives et spécialisées, mais ces propositions se sont heurtées au problème symétrique de la couverture des traitements et à la difficulté de formaliser l'ensemble des contraintes linguistiques nécessaires.

Face à ce constat, il est apparu incontournable et communément approuvé que des connaissances supplémentaires, c'est-à-dire non initialement disponibles dans les processus classiques de traitement, devaient être intégrées sous forme de jugements permettant d'attester la pertinence et la légitimité des hypothèses générées. Ainsi, pour chacune des tâches que comprend le TALN et qui potentiellement peut constituer ce que nous avons désigné sous le terme de cas d'indétermination, diverses sources de connaissances ont été envisagées pour contrôler la génération et surtout la propagation des hypothèses construites. On remarque alors que pour le contrôle d'un cas précis d'indétermination tel que la désambiguïsation du sens des mots, la sélection d'une meilleure analyse syntaxique ou le choix d'un attachement prépositionnel, plusieurs informations distinctives de natures variées peuvent être utilisées. Il peut en effet s'agir d'une exploitation de règles ou de contraintes linguistiques supplémentaires, d'une évaluation statistique de la fréquence des phénomènes linguistiques à l'aide d'un corpus, de calcul d'un score représentant un certain point de vue de jugement, etc. À travers quelques exemples (section 1.3), nous avons vu qu'exploité individuellement, le jugement apporté par chaque information ne permettait de traiter qu'une partie des indéterminations possibles. Différents travaux ont par la suite montré l'apport d'une prise en compte simultanée de plusieurs de ces sources de jugement afin d'obtenir un contrôle final plus fiable et robuste.

Notre étude de ce phénomène problématique nous a permis de constater que les différentes propositions de contrôle sont essentiellement consacrées à une tâche et un contexte d'application précis. Par rapport aux travaux existants, la démarche que nous proposons et que nous avons développée dans ce document constitue, d'après nos connaissances, l'unique tentative de formalisation de ce problème observé de manière récurrente au cours de l'application de procédures modulaires et séquentielles de traitement. En considérant l'apparition de cas d'indétermination comme un phénomène récurrent difficilement contournable intrinsèquement, nous rendons légitime l'usage de stratégies spécifiques de contrôle et montrons leur nécessité.

3.2 Une approche de contrôle décisionnelle basée sur de multiples critères de comparaison

Après une phase d'étude descriptive et explicative des cas d'indétermination, notre travail s'est focalisé sur l'identification d'une approche et d'une terminologie permettant de spécifier les propriétés du problème étudié. Lors de cette étape principalement théorique, nous nous sommes rapproché d'une hypothèse formulée par [Sabah, 1989] visant à considérer le traitement des cas d'indétermination comme un problème décisionnel. Une stratégie de contrôle s'apparente en effet à un processus rationnel d'évaluation d'alternatives, dont l'objectif est d'obtenir une recommandation de décision pouvant se matérialiser par un tri des hypothèses selon leur "degré" de pertinence, à un classement représentant leur pertinence respective ou finalement à une sélection d'un sous-ensemble d'hypothèses jugées comme les plus pertinentes. Quelle que soit la problématique visée, l'évaluation de la pertinence relative des différentes hypothèses concurrentes repose sur l'exploitation des différentes sources de connaissances supplémentaires. En apportant un jugement sur la pertinence de chacune des hypothèses concurrentes, ces connaissances ou informations distinctives apparaissent alors comme des critères de comparaison, nous rapprochant

ainsi un peu plus du paradigme décisionnel.

Cette description des propriétés communes aux différentes stratégies de contrôle envisagées jusqu'ici nous amène donc à formaliser ce processus comme un problème décisionnel basé sur plusieurs critères de comparaison. Devant l'hétérogénéité de ces critères et l'apparition de notions complexes liées à leur agrégation, comme la compensation, la contradiction ou la complémentarité, il semble indispensable d'exploiter une méthodologie adaptée et spécialisée dans l'étude de ces contextes décisionnels. C'est pour cette raison que nous avons proposé un rapprochement avec le domaine de l'Aide Multicritère à la Décision (AMCD), dont les travaux ont conduit au développement de méthodes spécialisées dans la construction de recommandations de décision à partir de multiples critères de comparaison. Parmi les différentes méthodes que nous avons étudiées (voir chapitre 3), les approches par surclassement et plus particulièrement les méthodes `ELECTRE` sont apparues comme les plus prometteuses, et ceci pour deux principales raisons. Tout d'abord, quelle que soit la problématique de recommandation visée (tri, classement ou sélection), ces méthodes s'appuient sur une notion commune de comparaison : le surclassement. Comme nous l'avons constaté lors du chapitre 4, cette propriété nous a notamment permis de développer une méthodologie générique de comparaison exploitable lors des trois problématiques de contrôle possible. Les méthodes d'AMCD par surclassement disposent également d'une propriété qui s'est révélée très importante dans notre contexte de TALN, celle de considérer l'expertise humaine comme une donnée importante du processus décisionnel. En effet, à travers un ensemble de paramètres préférentiels, un expert du domaine décisionnel peut influencer et même déterminer le comportement du processus décisionnel.

Cette prise en compte de l'expert lors de la mise en place d'une stratégie de contrôle est primordiale et constitue un apport et une avancée par rapport aux méthodes existantes. En effet, les principales approches de contrôle exploitent des méthodes statistiques pour observer sur corpus des traits discriminants entre les différentes hypothèses concurrentes. Les résultats de ces approches sont cependant énormément conditionnés par la pertinence des corpus d'apprentissage utilisés. De plus, pour de nombreux cas d'indétermination, de tels corpus d'apprentissage ne sont pas forcément disponibles. On constate également qu'un expert, linguiste ou informaticien, dispose souvent de connaissances suffisantes pour établir une stratégie de contrôle efficace. Ainsi, en héritant des travaux effectués en modélisation des préférences, les approches par surclassement définissent par un ensemble de paramètres préférentiels (poids, seuils de préférence, d'indifférence et de veto) les différentes connaissances qui permettent de déterminer le comportement d'un processus décisionnel. La mise en place d'une stratégie de contrôle ne repose donc plus sur la disponibilité d'un corpus d'apprentissage, mais peut être effectuée à partir de la formalisation des connaissances et intuitions d'un expert en tant que paramètres décisionnels.

Ce rapprochement avec l'AMCD complète donc la formalisation du contrôle des cas d'indétermination en tant que processus décisionnel. L'appropriation et l'adaptation des méthodes par surclassement nous ont permis de créer une intersection novatrice entre deux domaines de recherche, l'AMCD et le TALN, et de disposer d'une base méthodologique mature et spécialisée dans la résolution de problèmes tels que celui auquel nous sommes confronté.

3.3 Un système de contrôle opérationnel et fonctionnel

La première partie de nos travaux est principalement focalisée sur les aspects théoriques du contrôle des cas d'indétermination, en proposant notamment de formaliser cette tâche comme un

problème décisionnel basé sur l'exploitation de plusieurs critères de comparaison. En s'appuyant sur cette proposition, nous nous sommes ensuite focalisés sur la mise en place d'une méthodologie adaptée à cette formalisation. Cette démarche nous a conduit vers les méthodes par surclassement en AMCD qui proposent un cadre opérationnel et pragmatique pour la résolution de problèmes décisionnels permettant l'agrégation de critères hétérogènes.

Nous avons ensuite profité du contexte industriel dans lequel s'inscrivent nos travaux pour expérimenter et valider notre approche de contrôle. En effet, la chaîne de traitement `TiLT` constitue un support idéal d'application de notre approche. Lors d'une première phase très importante de conception, nous avons montré qu'il était possible de définir un module générique dédié au contrôle des différents modules de traitement qui composent `TiLT`. Ainsi, en nous appuyant sur une modélisation rigoureuse, nous avons pu développer un système complet facilitant la gestion et la manipulation des éléments décisionnels : hypothèses et critères, mais permettant également l'intégration de phases de contrôle complétant ainsi le processus initial d'analyse.

En reprenant la philosophie des approches par surclassement en AMCD, le système que nous avons conçu, implémenté et intégré au sein de l'architecture de `TiLT` place l'expert linguiste ou informaticien au cœur de la mise en place des stratégies de contrôle. Nous avons vu que cette caractéristique maintenait notre système utilisable pour le contrôle de cas d'indétermination pour lesquels aucun corpus d'apprentissage représentatif n'était disponible. Cependant, en cas de disponibilité d'un tel corpus, nous avons également proposé des extensions à notre système de contrôle initial. À l'aide de la méthode d'apprentissage de métriques `RELIEF` et de quelques heuristiques statistiques, nous assistons l'expert lors de la mise en place d'une stratégie de contrôle en lui suggérant des valeurs possibles pour les différents paramètres décisionnels. Pour compléter les fonctionnalités apportées par ce système de contrôle, nous avons également développé des outils connexes de visualisation et de manipulation des éléments décisionnels : configurations, critères, pré-ordres, tri, etc.

En cherchant à mettre à la disposition des développeurs de `TiLT` un système de contrôle complet, nous avons porté une attention particulière à la réutilisabilité et à la robustesse de nos outils. Bien que ces efforts pour rendre notre module opérationnel et fonctionnel dans une chaîne de traitement commerciale ne soient pas requis pour la préparation d'un doctorat, ils témoignent tout de même de notre volonté de donner une dimension industrielle à ces travaux. Afin que nos travaux se concrétisent par des outils fonctionnels au service des ingénieurs de recherche de l'équipe `Langues Naturelles`, nous nous sommes efforcé de concevoir et de développer une méthodologie conforme aux exigences d'un contexte industriel. Cette démarche et ces efforts nous ont notamment conduit à l'intégration de nos travaux dans la chaîne commerciale de traitement `TiLT` et à sa validation sur des cas concrets d'application.

De plus, en proposant des outils et documentations nécessaires à la mise en place de stratégies de contrôle, nous donnons la possibilité aux développeurs de `TiLT` d'exploiter ce système sur des cas d'indéterminations que nous n'avons pas traités. Les retours d'expérimentation en cours, notamment sur le contrôle du processus d'analyse sémantique, nous permettront à la fois de confirmer l'intérêt de nos travaux et de les faire progresser.

3.4 Quelques expérimentations parmi l'ensemble des cas de contrôle envisageables

À partir de ce travail de conception et d'intégration de notre système de contrôle au sein de l'architecture de traitement de `TiLT`, il est désormais possible d'envisager différents cadres d'expérimentation et d'évaluation du fonctionnement de la méthodologie et surtout de la pertinence des décisions émises. Parmi les différents processus d'analyse pouvant bénéficier de l'ajout d'une stratégie de contrôle, nous avons sélectionné trois cas différents d'indétermination nous permettant d'évaluer les trois problématiques prises en compte par notre approche de contrôle :

- le classement des rubriques d'indexation pour un service d'annuaire de professionnels ;
- le tri des paires d'expressions extraites d'un texte en fonction de l'existence ou non d'un lien de coréférence ;
- la sélection d'une meilleure transcription d'un SMS.

À travers ces trois cas d'indétermination, nous avons pu évaluer différents aspects de notre approche. Tout d'abord, nous avons constaté que la formalisation décisionnelle pouvait s'appliquer à ces différents cas d'indétermination.

Lors de l'étude du premier cas d'indétermination concernant le classement des rubriques d'indexation candidates pour une requête soumise à un annuaire de professionnels, nous avons mis en avant l'importance des critères utilisés lors de la comparaison des hypothèses. Cette expérimentation a suscité une réflexion intéressante sur la nature des sources de connaissances intégrées au processus d'analyse en tant que critères, et une prise de distance vis-à-vis des résultats que l'on pouvait attendre d'une stratégie de contrôle s'appuyant sur des critères non discriminants.
Lors de la seconde expérimentation, nous avons insisté sur l'intérêt d'une approche hybride basée à la fois sur des connaissances expertes définies *a priori* et sur des observations statistiques réalisées sur un corpus de référence représentatif.

La troisième expérimentation nous a principalement servi à l'évaluation de l'utilisabilité de notre système et son apport sur la simplification de l'intégration et de la manipulation de sources de connaissances supplémentaires. Cette seconde expérimentation nous a également permis de transformer un processus d'analyse purement symbolique en un système hybride, exploitant ainsi la complémentarité entre un système d'analyse symbolique et des mécanismes de contrôle statistique.

Pour chacune de ces expérimentations, nous avons vu que l'application de notre méthodologie de contrôle suscitait des perspectives d'amélioration intéressantes. Malgré la frustration que cela engendre, nous n'avons pas pu, dans le temps imparti pour ce projet, étudier toutes les facettes de contrôle envisageables sur chacun de ces cas d'indétermination. En effet, au lieu d'étudier de manière complète tous les problèmes soulevés par un de ces processus de traitement non déterministes, nous avons estimé important d'appliquer notre méthodologie sur des cas très variés d'indétermination, ceci afin de valider la généricité de notre système et son utilisabilité.

Un regard sur l'indéterminisme des chaînes de TALN

Quelle que soit la stratégie adoptée, les systèmes de TALN basés sur des représentations formelles de structures linguistiques se sont tous heurtés à la génération et la propagation d'hypothèses concurrentes et potentiellement erronées. Depuis fort longtemps, cet indéterminisme est considéré comme une limite majeure à l'essor des applications basées sur une analyse linguistique des données textuelles. Cependant, malgré le caractère récurrent de ce phénomène problématique, les travaux que nous avons menés constituent la première tentative de formalisation et de défini-

tion d'une approche de contrôle générique. Bien que comme tout choix méthodologique, le nôtre soit critiquable sur certains aspects, notre démarche conduit tout de même à une meilleure compréhension du problème étudié. De plus, la construction d'un système fonctionnel et opérationnel basé sur la formalisation proposée témoigne de notre volonté de mener ce projet informatique à terme et dans sa globalité. Nous avons donc suivi une démarche d'analyse complète d'un problème initial, reposant dans un premier temps sur une phase d'étude et de formalisation, puis sur la conception, la modélisation, l'implémentation et l'expérimentation d'une solution.

L'apport scientifique de notre travail se résume en deux points : tout d'abord l'analyse d'un problème récurrent et complexe à l'aide d'une démarche progressive traitant les différentes notions soulevées par le problème initial, puis la mise en place d'une intersection novatrice entre deux domaines de recherche jusqu'ici inconnus l'un de l'autre, afin notamment de bénéficier des apports mutuels entre les deux communautés concernées.

Nous n'avons évidemment pas la prétention de considérer nos travaux comme la solution à un problème aussi complexe que celui de la génération d'interprétations linguistiques concurrentes. Nous avons tout de même montré qu'il était primordial de considérer les systèmes de contrôle des hypothèses générées comme un composant légitime d'une chaîne de TALN. De plus, bien que nous nous soyons appuyé sur un exemple précis de processus d'analyse, celui de `TiLT`, la première partie de notre démarche descriptive offre une meilleure compréhension de l'indéterminisme des chaînes de traitement linguistique.

Ces travaux s'inscrivent donc dans un axe de recherche important de notre domaine, celui de l'évaluation des systèmes de TALN donnant notamment lieu à des campagnes d'évaluation (GRACE, EASY[Paroubek *et al.*, 2005], etc.). Bien que des résultats et des interprétations intéressants aient pu être obtenus, il reste encore de nombreux aspects théoriques, méthodologiques et pratiques à approfondir, chose que nous espérons et projetons de réaliser à court-terme.

A

Rappels méthodologiques

Cette annexe synthétise l'ensemble des notations et formules utilisées par les méthodes d'AMCD utilisées, à savoir `ELECTRE III` et `ELECTRE TRI`.

A.1 Formalisation commune des éléments décisionnels

Les différentes méthodes d'aide multicritère à la décision s'appuient sur deux éléments prépondérants :

- l'ensemble des hypothèses (également désignées sous les termes d'action, candidat ou solution) à comparer, que nous notons :

$$H : \{h_1, h_2, ..., h_n\}$$

- l'ensemble des critères utilisés pour comparer les hypothèses, que nous notons :

$$G : \{g_1, g_2, ..., g_m\}$$

où $g_k(h_i)$ correspond à la performance atteinte par l'hypothèse h_i sur le critère g_k. Ainsi, chaque hypothèse $h_i \in H$ est associée à un vecteur de performances :

$$h_i \rightarrow \{g_1(h_i), g_2(h_i), ..., g_m(h_i)\}$$

A.2 Formalisation des préférences

Le terme de préférence désigne à la fois les relations de comparaison que peuvent entretenir deux hypothèses et à la fois les connaissances associées aux critères déterminant leur modalité d'usage. La présentation de la formalisation adoptée de ces préférences différentie ces deux cas en utilisant les termes de préférences orientées "outputs" (pour la comparaison des hypothèses) et de préférences orientées "inputs" (sur les critères). D'une manière générale, une préférence exprime une connaissance du décideur soit sur la pertinence relative des hypothèses soit sur la nature d'un critère.

A.2.1 Préférences orientées "outputs"

Les approches par surclassement et plus généralement l'aide multicritère à la décision s'appuient sur les travaux issues des théories décisionnelles et de modélisation des préférences individuelles et collectives. Les préférences qualifiées d'orientées "outputs" constituent des situations

qu'un expert humain serait en mesure d'établir pour apprécier les différents degrés de qualité, de pertinence ou d'exactitude entre différents hypothèses concurrents.

La notion de surclassement a été introduite pour représenter une relation floue de préférence entre deux hypothèses concurrentes, mais permet tout de même d'établir les situations classiques de comparaison, à savoir la préférence, l'indifférence et l'incomparabilité.

- $h_i P h_j$ si $h_i S h_j$ et non $h_j S h_i$
- $h_i I h_j$ si $h_i S h_j$ et $h_j S h_i$
- $h_i R h_j$ si non $h_i S h_j$ et non $h_j S h_i$

A.2.2 Préférences orientées "inputs"

Également désignées sous le terme de paramètres préférentiels, ces préférences représentent les connaissances qu'un expert est en mesure de formuler sur les critères et la façon dont ils doivent être exploités lors du processus décisionnel. La méthode `ELECTRE III` exploite les préférences orientées "inputs" suivantes :

- un ensemble de seuils de préférence $P : \{p_1, p_2, ..., p_m\}$;
  p_k représente la plus petite différence $g_k(h_i) - g_k(h_j)$ traduisant une situation de préférence en faveur de h_i par rapport à h_j sur le critère g_k .
- un ensemble de seuils d'indifférence $Q : \{q_1, q_2, ..., q_m\}$;
  q_k représente la plus grande différence $g_k(h_i) - g_k(h_j)$ préservant une situation d'indifférence entre h_i et h_j sur le critère g_k .
- un ensemble de seuils véto $V : \{v_1, v_2, ..., v_m\}$.
  Où, v_k représente la plus petite différence $g_k(h_i) - g_k(h_j)$ tolérée pour que h_i et h_j restent comparable. Un tel seuil permet par exemple de considérer comme incomparable une hypothèse ayant une performance trop faible sur un critère pour qu'elle puisse être compensée par les autres critères.

A.3 Construction des relations de surclassement

Qu'il s'agisse de répondre à une problématique de tri, de classement ou de sélection, la comparaison de la pertinence relative des hypothèses concurrentes repose sur une notion commune de préférence, le surclassement. Une relation de surclassement entre deux hypothèses h_i et h_j, notée h_i S h_j, signifie qu'à partir des évaluations (performances de critères) dont on dispose sur ces deux hypothèses et des préférences orientées "inputs", h_i est au moins aussi pertinente que h_j.

En fonction de la précision des préférences orientées "inputs" introduites, différentes méthodes de construction de ces relations ont été proposées. Étant donné que certains de nos contexte de contrôle comportent des critères associés à des seuils de préférence, d'indifférence et de véto, nous exploitons la procédure de construction des relations de surclassement de la méthode `ELECTRE III`. Nous rappelons dans cette annexe les détails du calcul de l'indice de crédibilité $S(h_i, h_j) \in [0, 1]$ permettant d'établir ou non une relation de surclassement entre h_i et h_j.

$$C(h_i, h_j) = \frac{1}{P}.\sum_{k=1}^{m} w_k.c_k(h_i, h_j) \text{ où } P = \sum_{k=1}^{m} w_k$$

$$c_k(h_i,h_j)=\begin{cases}0, & \text{si } g_k(h_j)-g_k(h_i)\geq p_k\\ \frac{p_k-(g_k(h_i)-g_k(h_j))}{p_k-q_k}, & \text{si } q_k\leq g_k(h_j)-g_k(h_i)\leq p_k\\ 1, & \text{si } g_k(h_j)-g_k(h_i)\leq q_k\end{cases}$$

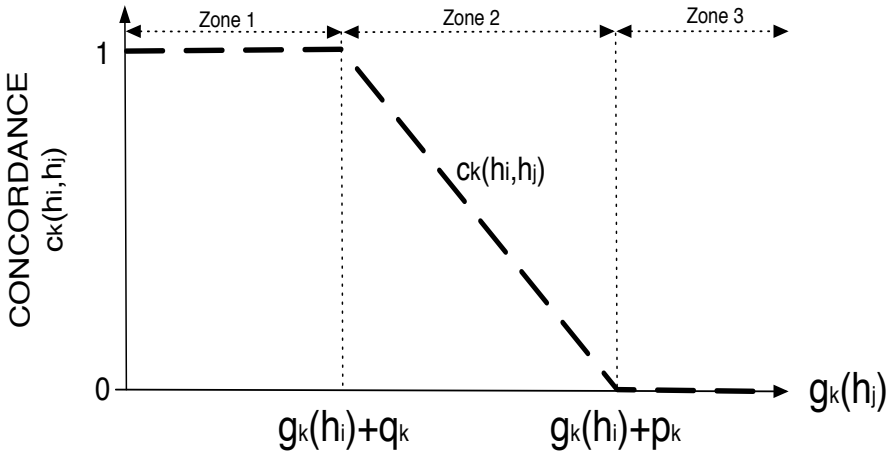

FIG. A.1 – Illustration graphique de la concordance

$$d_k(h_i,h_j)=\begin{cases}1, & \text{si } g_k(h_j)-g_k(h_i)\geq v_k\\ \frac{g_k(h_j)-g_k(h_i)-p_k}{v_k-p_k} & \text{si } p_k<g_k(h_j)-g_k(h_i)<v_k\\ 0, & \text{si } g_k(h_j)-g_k(h_i)\leq p_k\end{cases}$$

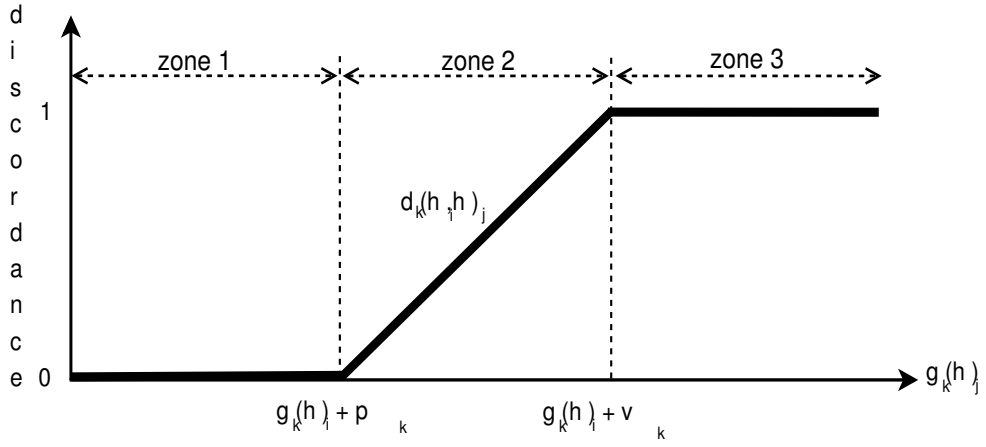

FIG. A.2 – Illustration graphique de la discordance

Soit F les indices des critères $G:\{g_1,g_2,...,g_m\}$ $(F:\{1,2,...,m\})$:

$$\sigma(h_i,h_j)=C(h_i,h_j)\prod_{k\in\overline{F}}\frac{1-d_k(h_i,h_j)}{1-C(h_i,h_j)}\text{où } \overline{F}=\{k\in F/d_k(h_i,h_j)>C(h_i,h_j)\}$$

$$h_i S h_j \text{ si } \sigma(h_i, h_j) \geq \alpha$$

où $\alpha \in [0.5, 1]$ représente un seuil final de coupe attestant que l'indice de crédibilité est suffisamment élevé pour que la relation de surclassement soit établie.

A.4 Compléments sur la méthode ELECTRE III : classement et sélection

Structure de préférences

Interprétation de la structure de préférences pour établir un classement des hypothèses

Initialement, la méthode ELECTRE III a été développée pour répondre à une problématique de classement des hypothèses concurrentes. Ce classement final correspondant à un pré-ordre partiel, est établie à partir de la confrontation de deux pré-ordres complets, l'un ascendant et l'autre descendant.

La construction de ces pré-ordres s'appuie sur les relations de surclassement valuées par leurs indices de crédibilité et d'un paramètre supplémentaire appelé seuil de discrimination. Ce seuil noté $s(\lambda)$ s'exprime en fonction des degré de crédibilité maximal noté λ.

$$s(\lambda) = \alpha + \beta.\lambda$$

où α et β ont été fixés dans notre système à des valeurs communément utilisées : $\alpha = 0,3$ et $\beta = -0,15$.

$s(\lambda)$ permet de différencier les relations de surclassement valables, c'est-à-dire dont l'indice de crédibilité est suffisamment significatif, et celle non valables, c'est-à-dire où $\sigma(h_i, h_j)$ n'est pas significativement plus élevé que $\sigma(h_j, h_i)$. Une relation est significative si :

$$\sigma(h_i, h_j) \geq \lambda - s(\lambda) \text{ et } |\sigma(h_i, h_j) - \sigma(h_j, h_i)| \leq s(\lambda)$$

Pour chacune des hypothèses h_i surclassante, on calcule son nombre de qualifications, noté $Q(h_i)$ correspondant au nombre d'hypothèses que h_i surclasse moins le nombre d'hypothèses qui surclassent h_i.

Les hypothèses de qualification maximale forment le premier ensemble d'hypothèse de rang 1 du classement et cette procédure sur les hypothèses restantes. Lorsque toutes les hypothèses ont été associées à un rang dans le classement on obtient un pré-ordre complet descendant.

L'obtention du pré-ordre complet ascendant repose sur le même algorithme, et diffère lors de la dernière où on ne conserve que les hypothèses ayant une faible qualification pour former les hypothèses de rang 1.

La construction du pré-ordre partiel final résulte du regroupement des pré-ordres complets ascendant et descendant. Les hiérarchies d'hypothèses similaires entre les deux pré-ordres antagonistes sont conservées pour former des relations de préférence. Les hypothèses appartenant au même rang dans les deux pré-ordres sont considérées comme indifférentes et les divergences forment des incomparabilités.

Interprétation de la structure de préférences pour extraire un sous-ensembles d'hypothèses préférables

Les relations construites par la méthode ELECTRE III ne sont pas initialement prévues pour être interprétées en vue de répondre à une problématique de sélection. Cependant, par rapport aux relations de surclassement générées par ELECTRE I ,qui est une méthode dédiée aux problèmes de sélection, les relations que nous utilisons diffèrent uniquement par l'existence d'un indice de crédibilité.

Bien que cette information n'est pas exploitée dans l'algorithme d'extraction d'un meilleur sous-ensemble d'hypothèses, la structure de préférences générées par la méthode ELECTRE III est utilisée pour répondre aux problématiques de rangement et de sélection.

Une structure de préférence correspond à un graphe orienté comprenant éventuellement des circuits. Afin de déterminer un meilleur sous-ensemble d'hypothèses, correspond au noyau du graphe, on procède tout d'abord à l'identification des circuits maximaux du graphe : Θ : $\{\theta_1, \theta_2, ..., \theta_n\}$.

Dans un second temps, des relations noté $\succ_\theta$ sont construites entre ces différents circuits selon la règle suivante :

$$\theta_i \succ_\theta \theta_j \Leftrightarrow [\theta_i \neq \theta_j \text{ et } [\exists h_b \in \theta_j, \exists h_a \in \theta_i \text{ tel que } h_a S h_b]]$$

La construction de cette relation asymétrique conduit à un graphe sans circuit admettant un unique noyau $N \subset \Theta$. Une méthode pour construire ce noyau est présentée dans [Roy and Bouyssou, 1993] page 367. On cherche les éléments de Θ, i.e. les circuits, tels qu'aucun autre circuit ne leur est préféré selon $\succ_\theta$. Ce premier ensemble est conservé et noté N^1. On supprime de Θ les éléments de N^1 et ceux dont un élément de N^1 lui est préféré. Tant que le sous-ensemble restant n'est pas vide, on réitère la procédure pour déterminer $N = N^1 \bigcup N^2 \bigcup ... \bigcup N^k$.

Pour déterminer le meilleur sous-ensemble d'hypothèses $H\prime$ à partir de N, on procède comme suit. Si N ne contient que des circuits d'une hypothèse, alors $H\prime = N$, sinon il faut déterminer dans chaque circuit composé de plusieurs hypothèses, un représentant. Le choix de ce représentant est quelque peu empirique, nous proposons de prendre celle surclassant le plus grand nombre d'hypothèses.

A.5 Compléments sur la méthode ELECTRE TRI : tri

La méthode ELECTRE TRI répond à la problématique de tri et procède donc à une comparaison, non pas des hypothèses entre elles, mais des hypothèses avec des profils d'acceptabilité des différentes classes.

L'affectation des hypothèses dans les différentes classes s'appuie sur un des deux algorithmes

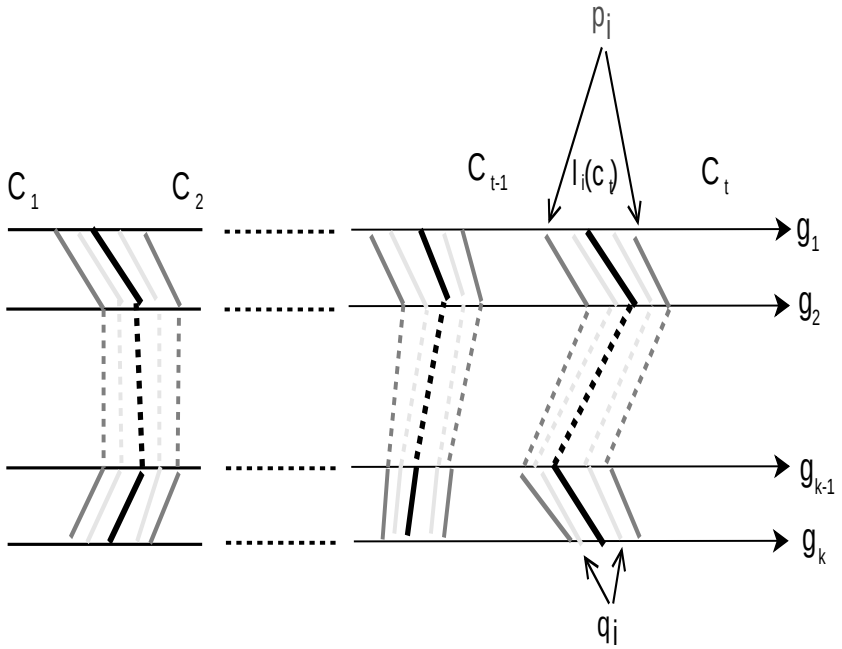

FIG. A.3 – Affectation des hypothèses aux classes définies *a priori* en utilisant la méthode `ELECTRE TRI`

suivants :

Précondition :
$C : \{c_t, c_{t-1}, ..., c_2, c_1\}$: classes définies *a priori*;
$H : \{h_1, h_2, ..., h_n\}$: hypothèses à trier;
Postcondition :
$\forall h_i \in H, \exists c_j \in C$ tel que $h_i \in c_j$;
Boucle :

```
for h_i ∈ H do
    for c_j ∈ {c_t, c_{t-1}, ..., c_2, c_1} do
        if h_i S c_j > α then
            h_i ∈ c_j
        end if
    end for
end for
```

Algorithme 4 : Algorithme de tri réalisé par `ELECTRE TRI` : version pessimiste

Précondition :
$C : \{c_1, c_2, ..., c_{t-1}, c_t\}$: classes définies *a priori*;
$H : \{h_1, h_2, ..., h_n\}$: hypothèses à trier;
Postcondition :
$\forall h_i \in H, \exists c_j \in C$ tel que $h_i \in c_j$;
Boucle :

for $h_i \in H$ **do**
  for $c_j \in \{c_t, c_{t-1}, ..., c_2, c_1\}$ **do**
    if $h_i S c_j > \alpha$ **then**
      $h_i \in c_j$
    end if
  end for
end for

Algorithme 5 : Algorithme de tri réalisé par `ELECTRE TRI` : version optimiste

Le résultat de l'application d'un de ces algorithmes offre une réponse directe à un problème de tri.

B

`Beslissing` : un module de contrôle décisionnel

B.1 Diagramme de classes détaillé du module de contrôle `Beslissing`

Les méthodes d'accès et de modification des variables de classe ne sont pas illustrées dans ce diagramme de classe afin de conserver une certaine lisibilité de l'ensemble des classes.

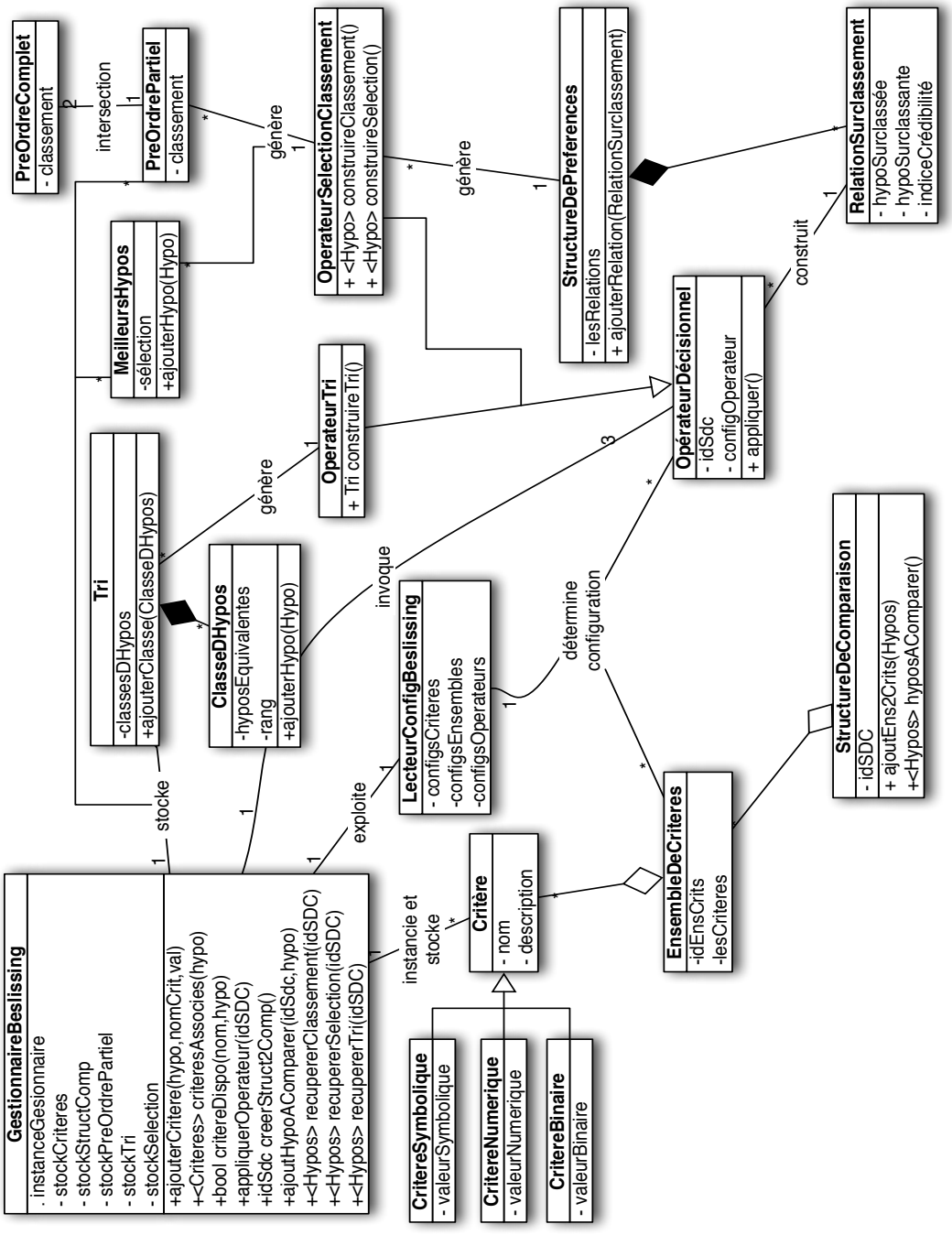

FIG. B.1 – Diagramme de classe du module décisionnel `Beslissing` intégré dans l'architecture de traitement `TiLT`

B.2 Fichier de configuration commenté du module de contrôle `Beslissing`

```
# ../ProfilSMS_Fige/configBeslissingControleChunkingSMS.ini
# Contrôle du processus de transcription : sélection des meilleurs terminaux, SVA et AGSN

[ConfigBeslissing]
Auteur = GregorySMITS
Date = 14092007
Description = sms

# CRITERES UTILISES
# SYNOPSIS : CriteresDisponibles=(NomDuCritere,Type:Numerique|Binaire|Symbolique,
#valeurParDefaut)*

CriteresDisponibles=(CoutTerminal,Numerique,_);(TermDistCor,Numerique,_);
(frequenceLexicale,Numerique,_);(scoreControleSemantique,Numerique,_);
(scoreControleSyntaxique,Numerique,_);(rangControleSemantique,Numerique,_);
(rangControleSyntaxique,Numerique,_);(TermAppartientSolBigrammes,Binaire,_);
(GS1appartientSolBigrammes,Binaire,_);(SVAscoreSolBigrammes,Numerique,_);
(AGSNscoreSolBigrammes,Numerique,_);(AGSNvalideSelonTri,Binaire,_);
(AGSNvalideSelonGram,Binaire,_);(freqLexForme,Numerique,_)


#ENSEMBLES DE CRITERES UTILISES PAR LES OPERATEURS
# SYNOPSIS : EnsemblesDeCriteresDisponibles = (NomDeLEnsemble,
[NomCritere:Poids:SeuilPref:SeuilIndif:SeuilVeto:LimiteAcceptabilite:
OperateurAgg:Moy|Max|Min|Som]*)*
# LimiteAcceptabilite définit uniquement lorsque l'ensemble est utilisé par un opérateur de
# OperateurAgg utilisé uniquement par les opérateurs d'agrégation

EnsemblesDeCriteresDisponibles=(criteresPreTerminal,frequenceLexicale:0.1:_:_:_:_:_
freqLexForme:0.1:_:_:_:_:_  scoreControleSemantique:0.2:_:_:_:_:_
rangControleSyntaxique:0.1:_:_:_:_:_ rangControleSemantique:0.1:_:_:_:_:_
 scoreControleSyntaxique:0.2:_:_:_:_:_ );(criteresTerminal,TermDistCor:0.3:-2:-1:-3:_:_
  frequenceLexicale:0.1:20:10:50:_:_  freqLexForme:0.3:5:10:20:_:_
  rangControleSyntaxique:0.3:2:1:4:_:_ TermAppartientSolBigrammes:0.3:_:_:_:_:_);
  (criteresGS1,frequenceLexicale:0.1:_:_:_:_:Moy freqLexForme:0.1:_:_:_:_:Moy
  rangControleSyntaxique:0.1:_:_:_:_:_   Max scoreControleSyntaxique:0.2:_:_:_:_:Min);
  (criteresSVA,SVAscoreSolBigrammes:0.3:0.2:0.1:_:_:_   frequenceLexicale:0.2:20:10:50:_:_
  freqLexForme:0.2:5:5:10:_:_ rangControleSyntaxique:0.1:2:1:4:_:_
  scoreControleSyntaxique:0.2:30:15:200:_:_);
  (criteresTriAGSN,AGSNscoreSolBigrammes:1.0:0.2:0.1:_:_:_);
  (criteresAGSN,AGSNscoreSolBigrammes:0.5:0.2:0.1:_:_:_ frequenceLexicale:0.2:20:10:50:_:_
  freqLexForme:0.2:5:5:10:_:_ rangControleSyntaxique:0.1:2:1:4:_:_
  scoreControleSyntaxique:0.2:30:15:200:_:_)
```

```
# OPERATEUR UTILISE
# SYNOPSIS : OperateursDecisionnelsDisponibles=(NomOperateurClassementSelection,
NomDeLEnsemble) | (NomOperateurTri,NomDeLEnsemble,DefinitionProfilsAcceptabilite,
coupeAlpha)

OperateursDecisionnelsDisponibles = (opClassementTerminaux,criteresTerminal);
(opAggAGSNGS1,criteresGS1);(opAggSVAGS1,criteresGS1);
(opClassementSVA,criteresSVA);(opTriAGSN,criteresTriAGSN,0.8,0.8);
(opClassementAGSN,criteresAGSN)
```

C

Outils connexes au module décisionnel `Beslissing`

C.1 Interface de configuration du module décisionnel `Beslissing`

FIG. C.1 – Interface graphique de configuration et de définition des modèles de préférences : exemple de définition des critères à utiliser

FIG. C.2 – Interface graphique de configuration et de définition des modèles de préférences : exemple de définition d'une opération de contrôle

C.2 `CorpusTagger` : une interface d'annotation des hypothèses de référence et de construction des tables de performances

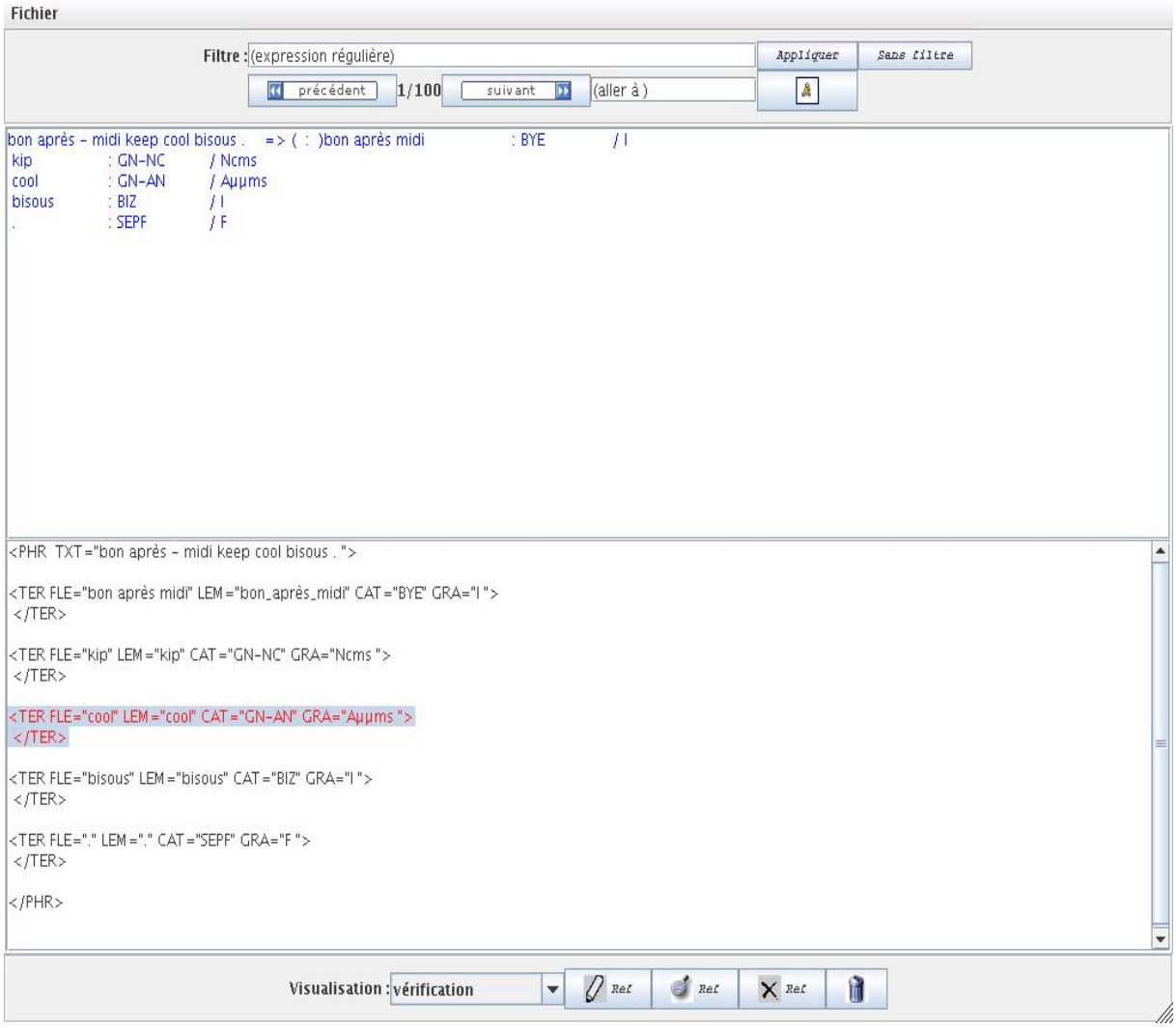

FIG. C.3 – Interface graphique de construction d'un corpus de références et de la table de performances associée à partir des hypothèses concurrentes générées par un module de traitement de `TiLT` : application de vue (transformation `XSLT` sur le fichier `XML` des sorties) pour déterminer plus facilement leur validité

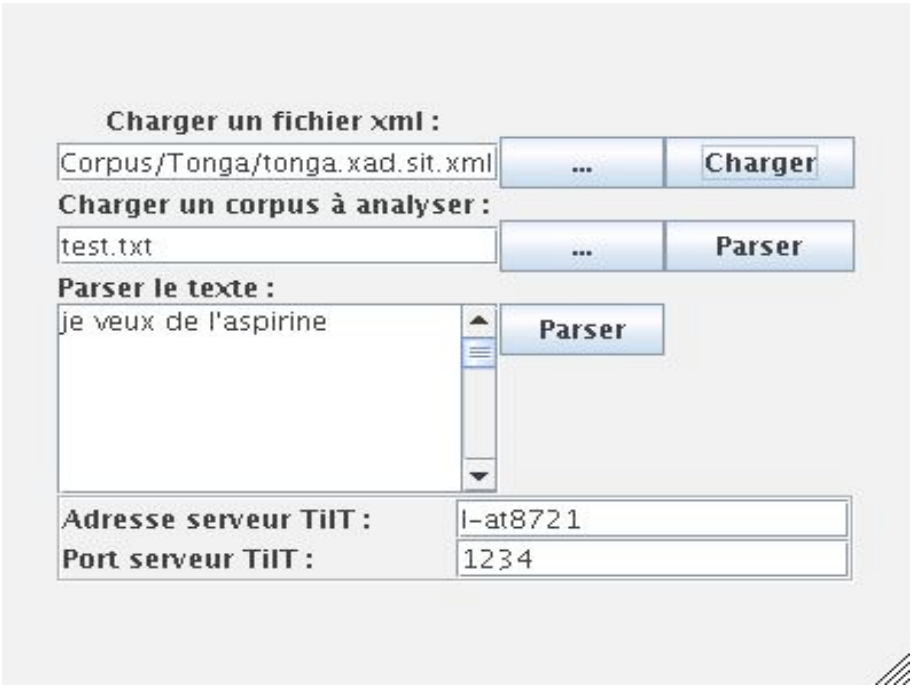

FIG. C.4 – Interface graphique de construction d'un corpus de références et de la table de performances associée à partir des hypothèses concurrentes générées par un module de traitement de `TiLT` : chargement d'un fichier `XML` des sorties d'un module ou analyse directe d'une phrase

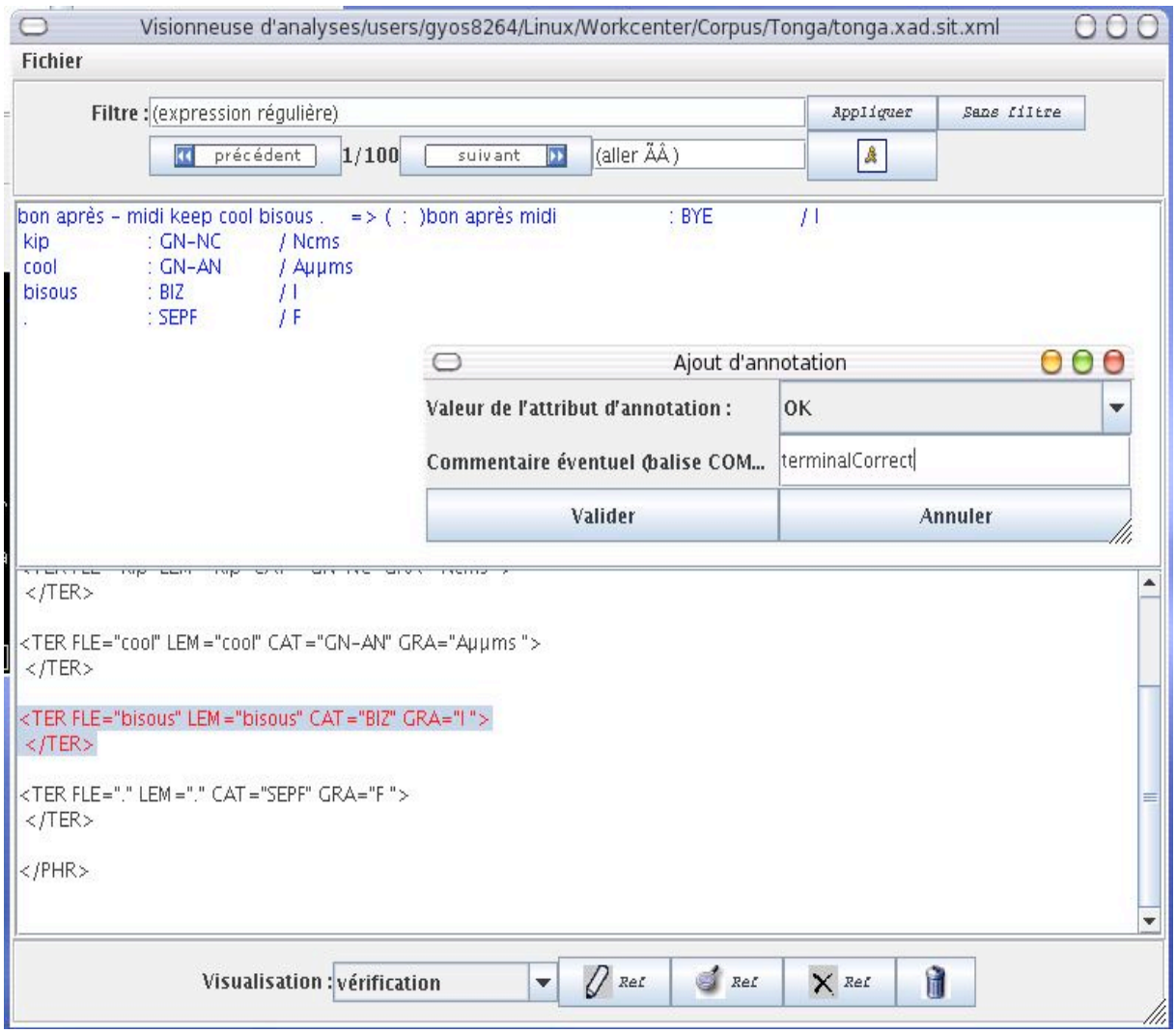

FIG. C.5 – Interface graphique de construction d'un corpus de références et de la table de performances associée à partir des hypothèses concurrentes générées par un module de traitement de `TiLT` : intégration d'annotations dans le fichier `XML` des sorties

D

Expérimentation autour de la résolution de la coréférence

D.1 Extraits du corpus “Le Monde” utilisé pour la résolution des liens de coréférence

Voici un paragraphe extrait du corpus :

" L’ épopée espagnole de Mr Silvio Berlusconi commence mal . Moins de deux mois avant la date du 5 mars , prévue pour le début des émissions , le torchon brûle , en effet , entre les trois principaux actionnaires de la chaine de télévision privée Gestavision - Telecinco , qui détiennent , chacun , 25% des parts : le groupe Fininvest , du magnat italien , la maison d’ édition Anaya et l’ Organisation nationale des aveugles d’ Espagne , la ONCE . Avec l’ appui d’ un allié minoritaire , les représentants de Mr Berlusconi et de la ONCE ont forcé , le jeudi 11 janvier , lors d’ une assemblée générale tumultueuse , la destitution du président et de l’ administrateur délégué de la chaine , MM German Sanchez et Pedro Higuera , tous deux du groupe Anaya , et leur exclusion du conseil d’ administration . Leurs remplaçants seront nommés lors d’ une prochaine assemblée générale . Cette rupture virtuelle fait suite à plusieurs semaines de polémiques publiques entre Mr Berlusconi et la maison d’ édition espagnole . Celle - ci avait ouvert le feu , en affirmant que l’ homme d’ affaires italien cherchait à contrôler seul le projet . Plus concrètement , les représentants d’ Anaya l’ ont accusé d’ avoir mis sur pied un habile montage financier , permettant à ses propres sociétés d’ assurer en régime de monopole les activités les plus lucratives de la chaine . Ainsi , la gestion publicitaire était confiée à une filiale de Fininvest , Publiespana , tandis que la production de programmes revenait à une autre société contrôlée par l’ homme d’ affaires italien , Videotime . Autant de décisions , souligne le groupe Anaya , qui n’ ont jamais obtenu l’ accord de l’ ensemble des associés et qui donnent à penser que le magnat italien confond ses propres intérêts avec ceux de Gestavision . Du côté de Mr Berlusconi , on affirme au contraire que , lors d’ une réunion tenue le 5 mai 1989 , avant même l’ attribution des trois chaines privées par le gouvernement espagnol , les trois grands actionnaires de Gestavision s’ étaient mis d’ accord sur un tel schéma . On fait aussi valoir que la société Fininvest avait , dès le départ , engagé seule , à ses risques et périls , les coûteuses dépenses (quelque 20 millions de dollars) permettant d’ équiper techniquement la chaine . (...) "

Ci-dessous l'analyse corrigée générée sur la phrase "L' épopée espagnole de Mr Silvio Berlusconi commence mal . " par TiLT, complétée par une annotation des expressions et de leurs éventuels liens de coréférence :

```
<?xml version="1.0"  encoding="iso-8859-1" ?>
<TLT>
    <PARA COD="ISO-8859-1" FOR="Fr" ISO="fr" LAN="francais">
    <PHR TXT="L' épopée espagnole de Mr Silvio Berlusconi commence mal . ">
     <EXPRESSION IDREF="0" NPR="0" STX="DDEF" TYP="AUTR">
<TER CAT="GN-D" FLE="L'" FONC="DET" ID="0" LEM="le" PI="1" REGLE="DET-1"
TRA="( GENRE/FEMININ NOMBRE/SINGULIER COUTTERM/2 CATEP/GN-D )"/>
<TER CAT="GN-NC" FLE="épopée" FONC="SUJ-V" ID="1" LEM="épopée" PI="7"
REGLE="GV-A1" TRA="( SY_SEM_DEF/ SY_OTY_DETER/ SY_TYP_CIRC/+ COUTTERM/5
CATEP/GN-NC GENRE/FEMININ NOMBRE/SINGULIER SY_TYP_QUAL/ )"/>
<TER CAT="GN-AN" FLE="espagnole" FONC="MOD-N" ID="2" LEM="espagnol" PI="1"
REGLE="EPITD-1-1" TRA="( GENRE/FEMININ NOMBRE/SINGULIER COUTTERM/3
CATEP/GN-AN ADJSEM/GEOG )"/>
</EXPRESSION>
<TER CAT="GP-S" FLE="de" FONC="PREP" ID="3" LEM="de" PI="6" REGLE="PREP-40"
TRA="( COUTTERM/1 INTRO_CAS/INDEF CATEP/GP-S PREP_FORME/DE )"/>
<EXPRESSION IDREF="1" NPR="1" STX="SDPR" TYP="PERS">
<TER CAT="GN-NP" FLE="Mr" FONC="NOM+" ID="4" LEM="Mr" PI="5" REGLE="PROP-2"
TRA="( COR/NPR COUTTERM/4 CATEP/GN-NP )"/><TER CAT="GN-NP" FLE="Silvio"
  FONC="NOM+" ID="5" LEM="Silvio" PI="6" REGLE="PROP-2" TRA="( COR/NPR COUTTERM/4 CATEP/GN-NP )"/>
  <TER CAT="GN-NP" FLE="Berlusconi" FONC="CIRC" ID="6" LEM="Berlusconi" PI="1"
  REGLE="CIRC-GENERIQUE-NC" TRA="( COR/NPR INTRO_CAS/INDEF COUTTERM/4
  PREP_FORME/DE CATEP/GN-NP CATEPF/GP-NP )"/>
  </EXPRESSION>
 <TER CAT="GV-PT" FLE="commence" ID="7" LEM="commencer" TRA="( SY_OTY_OBJ_GENRE/MASCULIN
   TRANSITIF/OUI PERSONNE/3PRS/1PRS SY_OTY_OBJ_NOMBRE/SINGULIER SY_OTY_OBJD/+
    TEMPS/PRESENT COUTTERM/4 MODE/SUBJ/IND CATEP/GV-PT CATEPF/PHRASE NOMBRE/SINGULIER SY_OTY_SUJ/+ )"/>
       <TER CAT="GN-NC" FLE="mal" FONC="COD-V" ID="8" LEM="mal" PI="7" REGLE="OBJD-30"
    TRA="( GENRE/MASCULIN NOMBRE/SINGULIER COUTTERM/5 DET_PART/OUI CATEP/GN-NC )"/>
   <TER CAT="SEPF" FLE="." FONC="SEPPH" ID="9" LEM="." PI="7" REGLE="SEPF-1"
    TRA="( PONCTUATION/ CATEP/SEPF )"/>
    </PHR>
    (...)
    </PARA>
</TLT>
```

D.2 Critères d'évaluation des couples candidats antécédent/reprise

Soit $P(e_1, e_2)$ une paire d'expressions où e_1 correspond à l'antécédent candidat et e_2 à la reprise candidate :

TAB. D.1 – Critères disponibles et exploitées pour l'identification des cas de coréférence

Nom du critère	**Type**	**Description**
isClosest	Binaire	e_1 est l'expression la plus proche précédant e_2
countOcc	Numérique	Nombre d'occurrences de e_1 dans le texte
distExp	Numérique	Nombre d'expressions nominales entre e_1 et e_2
distSent	Numérique	Nombre de phrases entre e_1 et e_2
distTer	Numérique	Nombre de mots entre e_1 et e_2
distParag	Numérique	Nombre de paragraphes entre e_1 et e_2
sameSentence	Binaire	Vrai si les deux expressions appartiennent à la même phrase
isDefinite	Binaire	Vrai si e_2 est précédée par un déterminant défini
isDemonst	Binaire	Vrai si e_2 est précédée par un déterminant démonstratif ou un pronom démonstratif
isPossessive	Binaire	Vrai si e_2 est précédé par un déterminant possessif ou un pronom possessif
isBareNP	Binaire	Vrai si e_2 n'est pas précédé par un déterminant
isIndef	Binaire	Vrai si e_2 est précédé par un déterminant indéfini
isSubj	Binaire	Vrai si e_1 a une fonction de sujet dans la phrase
tightParallel	Binaire	Vrai si les deux expressions ont une fonction de sujet dans des phrases adjacentes
looseParallel	Binaire	Vrai si les deux expressions ont une fonction de sujet
agrNum	Binaire	Vrai si les deux expressions n'ont pas de marques de nombre contradictoires
agrNbrStrict	Binaire	Vrai si les deux expressions sont de même nombre
agrGen	Binaire	Vrai si les deux expressions n'ont pas de marques de genre contradictoires
agrGenStrict	Binaire	Vrai si les deux expressions sont de même genre
attributive	Binaire	Vrai si les deux expressions apparaissent dans une construction copulative, i.e. sont des arguments d'un verbe d'état
appo	Binaire	Vrai si les deux expressions apparaissent dans une apposition
wordsInCommon	Numérique	Pourcentage de mots en commun entre les deux expressions
sameArgDomain	Binaire,null	Vrai si les deux expressions dépendent du même verbe (même structure argumentative du verbe) ou de la même préposition ; Null en cas de manque d'information
isAcronym	Binaire	Vrai si une expression est l'acronyme de l'autre
strSimil	Numérique	Distance typographique de Levenshtein, correspond au coût de la suite de transformations élémentaires la moins coûteuse pour transformer e_1 en e_2
subString	Binaire	Vrai si une expression est une sous-chaîne de l'autre
similModif	Binaire	Vrai si les deux expressions ont au moins un modifieur en commun

D.3 Modèles de préférences

Les critères disponibles illustrées dans l'annexe D.2 qui ne sont pas présents dans les modèles de préférences correspondent à des critères considérées comme non pertinents (discriminant) pour le type de paires concerné. Cette décision émane soit de l'expert soit d'une interprétation des résultats de la méthode `RELIEF` (poids inférieur à $0,01$).

D.3.1 Modèles de préférences définis *a priori* par un expert

`NPR-NPR`

TAB. D.2 – Modèle de préférences expert pour les paires `NPR-NPR`

Rang	**Nom du critère**	**Poids** $[0,10]$	**Seuil préf.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	strSimil	4	0,2	0,1	0,27	0,3
2	subString	3	-	-	-	1
3	wordsInCommon	1	0,1	0,05	0,25	0,15
4	isAcronym	1	-	-	-	1
5	countOcc	0,5	2	1	-	3
6	isSubj	0,5	-	-	-	1
7	tightParallel	0,5	-	-	-	1
8	looseParallel	0,5	-	-	-	1
9	sameArgDomain	0,5	0,1	0,1	0,5	1
10	appo	0,5	0,1	0,1	0,5	1
11	agrNum	0,5	-	-	-	1
12	agrGen	0,5	-	-	-	1

Le seuil de coupe établi par l'expert pour ce type de paires est de $\alpha = 0,7$.

`NPR-NCOM`

TAB. D.3 – Modèle de préférences expert pour les paires `NPR-NCOM`

Rang	**Nom du critère**	**Poids** $[0, 10]$	**Seuil préf.**	**Seuil indiff**	**Seuil veto**	**Limite**
1	countOcc	1,5	3	2	-	2
2	appo	1,5	-	-	-	1
3	isClosestCandidate	1	-	-		1
4	distExp	1	0,15	0,05	-	0,55
5	distSent	1	0,1	0,05	-	0,65
6	distTer	1	0,1	0,05	-	0,5
7	distParag	1	0,1	0,05	-	0,7
8	tightParallel	0,8	-	-	-	1
9	isSubj	0,5	-	-	-	1
10	looseParallel	0,5	-	-	-	1
11	isBareNP	0,25	0,0	0,0	0,5	1
12	isIndef	0,25	0,0	0,0	0,01	1
13	agrNum	0,25	-	-	-	1
14	agrGen	0,25	-	-	-	1
15	attributive	0,25	-	-	-	1

Le seuil de coupe établi par l'expert pour ce type de paires est de $\alpha = 0,5$.

`NPR-PRON`

TAB. D.4 – Modèle de préférences expert pour les paires `NPR-PRON`

Rang	**Nom du critère**	**Poids** $[0, 10]$	**Seuil préf.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	isClosestCandidate	2	-	-	-	1
2	countOcc	2	2	1	-	1
3	distExp	1	0,15	0,1	0,5	0,7
4	distSent	1	0,1	0,05	0,3	0,85
5	distTer	1	0,15	0,1	0,5	0,7
6	distParag	1	0,05	0,02	0,2	0,9
7	isDefinite	0,5	-	-	-	1
8	isDemonst	0,5	-	-	-	1
9	isSubj	0,5	-	-	-	1
10	sameArgDomain	0,25	0,0	0,0	0,5	1
11	tightParallel	0,25	-	-	-	1
12	looseParallel	0,25	-	-	-	1
13	isIndef	0,25	0,0	0,0	0,5	1
14	agrNum	0,25	-	-	-	1
15	agrGen	0,25	-	-	-	1
16	sameSentence	0,25	-	-	-	1

Le seuil de coupe établi par l'expert pour ce type de paires est de $\alpha = 0,6$.

D.3.2 Modèles de préférences suggérés

`NPR-NPR`

TAB. D.5 – Modèle de préférences suggéré à partir des tables de performances dédiées à l'apprentissage pour les paires `NPR-NPR`

Rang	**Nom du critère**	**Poids** $[0,1]$	**Seuil préf.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	wordsInCommon	0.25	0.2	0.0	0.0005	0,21
2	subString	0.25	-	-	-	1
3	isSimil	0.2	-	-	-	1
4	strSimil	0.2	0.26	0.12	0.54	0.71
5	countOcc	0.03	2	1	-	2
8	isSubj	0.03	-	-	-	1
9	tightParallel	0.01	-	-	-	1
10	looseParallel	0.01	-	-	-	1
11	agrNum	0.01	-	-	-	1
12	agrGen	0.01	-	-	-	1

Le seuil de coupe est également suggéré automatiquement à partir des tables de performances. L'indice de crédibilité de surclassement est calculé entre chaque hypothèse contenue dans les tables de performances et le profil d'acceptabilité de la classe des hypothèses valides. La valeur suggérée pour le seuil de coupe $alpha \in [0,1]$ correspond à la limite permettant de discriminer avec un taux d'erreur minimal les hypothèses valides des hypothèses non valides.

Pour les paires `NPR-NPR`, ce seuil a été établi à $0,75$.

`NPR-NCOM`

TAB. D.6 – Modèle de préférences suggéré à partir des tables de performances dédiées à l'apprentissage pour les paires `NPR-NCOM`

Rang	**Nom du critère**	**Poids** $[0,1]$	**Seuil préf.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	agrGen	0.28	-	-	-	1
2	agrNum	0.23	-	-	-	1
3	isDefinite	0.15	-	-	-	1
4	distExp	0.08	0.18	0.11	-	0.22
5	distTer	0.07	0.11	0.06	-	0.09
6	distSent	0.05	0,08	0,05	-	0,3
7	isSubj	0.05	-	-	-	1
8	tightParallel	0.02	-	-	-	1
9	looseParallel	0.02	-	-	-	1
10	agrNumStrict	0.02	-	-	-	1
11	agrGenStrict	0.01	-	-	-	1
12	wordsInCommon	0.01	-	-	-	1
13	sameSentence	0.01	-	-	-	1

Le seuil de coupe α utilisé pour les paires `NPR-NCOM` est de $0,541$.

`NPR-PRON`

TAB. D.7 – Modèle de préférences suggéré à partir des tables de performances dédiées à l'apprentissage pour les paires `NPR-PRON`

Rang	**Nom du critère**	**Poids** $[0,1]$	**Seuil préf.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	distSent	0.23	0.21	0.08	0.45	0.51
2	distExp	0.19	0.15	0.12	0,25	0.3
3	isClosestCandidate	0.15	-	-	-	1
4	isSubj	0.12	-	-	-	1
5	countOcc	0.06	1	0	-	1
6	agrGen	0.06	-	-	-	1
7	distTer	0.05	0.06	0.03	-	0.06
8	agrGenStrict	0.05	-	-	-	1
9	tightParallel	0.03	-	-	-	1
10	looseParallel	0.02	-	-	-	1
11	sameSentence	0.02	-	-	-	1
12	agrNum	0.01	-	-	-	1
13	agrNbrStrict	0.01	-	-	-	1

Le seuil de coupe α utilisé pour les paires `NPR-NCOM` est de $0,551$.

Exemple de courbes de distributions des paires du corpus d'apprentissage pour l'identification des zones de préférence, d'indifférence et de veto

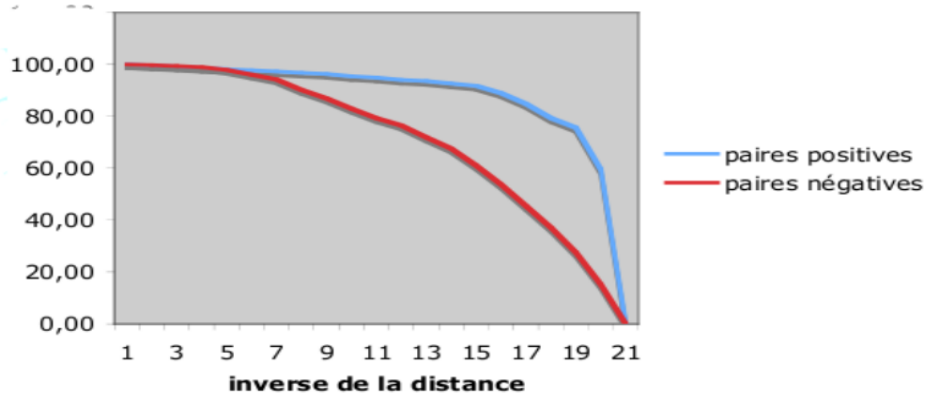

FIG. D.1 – Courbes de distribution construites à partir du corpus d'apprentissage pour les paires NPR-NPR sur le critère de similarité typographique

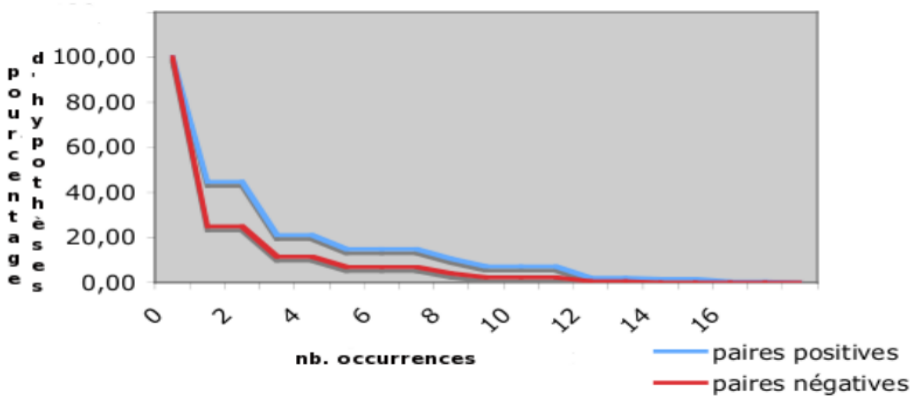

FIG. D.2 – Courbes de distribution construites à partir du corpus d'apprentissage pour les paires NPR-NCOM sur le critère du nombre d'occurrences de l'antécédent

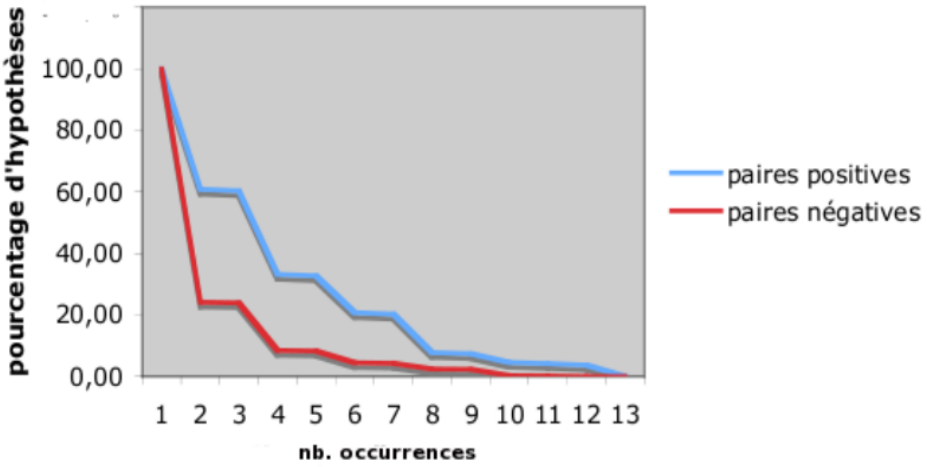

FIG. D.3 – Courbes de distribution construites à partir du corpus d'apprentissage pour les paires NPR-PRON sur le critère du nombre d'occurrences de l'antécédent

D.3.3 Modèles de préférences mixtes

`NPR-NPR`

Tab. D.8 – Modèle de préférences mixte, établi par l'expert à partir du modèle de préférences suggéré pour les paires `NPR-NPR`

Rang	**Nom du critère**	**Poids** $[0,1]$	**Seuil préf.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	wordsInCommon	0.25	0.2	0.0	0.0005	0,21
2	subString	0.25	-	-	-	1
3	isSimil	0.2	-	-	-	1
4	strSimil	0.2	0.26	0.12	0.54	0.71
5	isAcronym	1	-	-	-	1
6	sameArgDomain	0,5	0,1	0,1	0,5	1
7	appo	0,5	0,1	0,1	0,5	1
8	countOcc	0.03	2	1	-	2
9	isSubj	0.03	-	-	-	1
10	tightParallel	0.01	-	-	-	1
11	looseParallel	0.01	-	-	-	1
12	agrNum	0.01	-	-	-	1
13	agrGen	0.01	-	-	-	1

Comme pour les modèles de préférences suggérés, le seuil de coupe α est établi automatiquement à partir des indices de crédibilité de surclassement calculés sur chaque hypothèse contenue dans les tables de performances. Dans ce cas précis des paires `NPR-NPR`, le seuil de coupe α utilisé est de $0,77$.

`NPR-NCOM`

TAB. D.9 – Modèle de préférences mixte, établi par l'expert à partir du modèle de préférences suggéré pour les paires `NPR-NCOM`

Rang	**Nom du critère**	**Poids** $[0,1]$	**Seuil préf.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	agrGen	0.19	-	-	-	1
2	agrNum	0.15	-	-	-	1
3	appo	0.09	-	-	-	1
7	isDefinite	0.09	-	-	-	1
4	attributive	0,08	-	-	-	1
5	isIndef	0,07	0,0	0,0	0,01	1
6	isBareNP	0,07	0,0	0,0	0,5	1
8	distExp	0.07	0.18	0.11	-	0.22
9	distTer	0.05	0.11	0.06	-	0.09
10	distSent	0.04	0,08	0,05	-	0,3
11	isSubj	0.04	-	-	-	1
12	tightParallel	0.01	-	-	-	1
13	looseParallel	0.01	-	-	-	1
14	agrNumStrict	0.01	-	-	-	1
15	agrGenStrict	0.01	-	-	-	1
16	wordsInCommon	0.01	-	-	-	1
17	sameSentence	0.01	-	-	-	1

Le seuil de coupe α utilisé pour les paires `NPR-NCOM` est de $0,511$.

`NPR-PRON`

TAB. D.10 – Modèle de préférences mixte, établi par l'expert à partir du modèle de préférences suggéré pour les paires `NPR-PRON`

Rang	**Nom du critère**	**Poids**	**Seuil préf.**	**Seuil indiff.**	**Seuil veto**	**Limite**
1	distSent	0.18	0.21	0.08	0.45	0.51
2	distExp	0.15	0.15	0.12	0,25	0.3
3	isClosestCandidate	0.12	-	-	-	1
4	isSubj	0.09	-	-	-	1
5	countOcc	0.05	1	0	-	1
6	agrGen	0.05	-	-	-	1
7	distTer	0.04	0.06	0.03	-	0.06
8	agrGenStrict	0.04	-	-	-	1
9	sameSentence	0,04	-	-	-	1
10	isDefinite	0,04	-	-	-	1
11	isDemonst	0,04	-	-	-	1
12	sameArgDomain	0,04	0,0	0,0	0,5	1
13	isIndef	0,04	0,0	0,0	0,5	1
14	tightParallel	0.02	-	-	-	1
15	looseParallel	0.02	-	-	-	1
16	sameSentence	0.02	-	-	-	1
17	agrNum	0.01	-	-	-	1
18	agrNbrStrict	0.01	-	-	-	1

Le seuil de coupe α utilisé pour les paires `NPR-PRON` est de $0,537$.

E

Expérimentation autour de la transcription de SMS

E.1 Extraits du corpus utilisé pour la transcription de SMS

Corpus de `Louvain`

- OK, les pirates n'ont ka bien se tenir. Bon voyage ! Math et Manu
- Ici c la nuit kil fait beau. Ciel étoilé. Je te baise les lèvres. M
- Sans blem . Dormez bien, oiseaux du paradis... Sans jeu de mots. Bisous
- Troupinette,on est le 15 et la mission est annulée,jveux d'abord ton avis..Fais de beaux reves,miss Ensapinée,moi je plonge dans les miens !A très vite ! !bisouse
- T'me fé bp2b,j'ss happy2t'avoir&passé du tps ac3.On va dodo en m tps alor,c mieu k'rien m si j'aurai voulu t'avoir pré2moi,c p-e tro bo co reve car lé exam approche.Dor b mon chéri,bonn nui&fé2bo rev2moi mon p'ti ange.j'croi k'je ss vraimen trè trè amoureuse2 3,j'sai pa fer autremen é2t'résisté,c tro dur
- ..j'allais rouler 1pelle à Morphée ! ! jrigol ! Tt ok 2 mon coté.(bcp 2 filles mais ttes oqp ! !)En th. on devrait se voir vers noel. Et en janvier chu à Evere.. A+
- Ds ton tps de midi de demain tu saurais aller au disport pr donner les dates ?Sinon dis le moi et j'irai en revenant de bxl.Bonne nuit.Gros bisous
- jsspamalereu,jssamoureu.jné jms rsenti kch comsa av2tconètr.moiO6Gtroenvi.ttskGenvi2tdir, Cds lcreu2tonOreil kjve legliC.jarivpa amCparé2limag2toiémoi l1 ds lotr
- Fait vite on attend plus que toi
- On fait comme on a dit alors ? A demain ! Bonne nuit ! Guibert
- Jte prévien ce soir chui en forme
- Bonne soirée,on se rejoint en reve ? T'as passé 1 bonne journée ? Ici tout va bien ! Je pense bien fort à toi ! Tu as eu Flo au tél ? GROS BISOUS DE TENDRESSE...
- En plus la photo que tu ma envoyer cet après midi ma mis l'eau à la bouche.....
- J't'en veu pa mè c pa cool à entendr c'genr d'réflexion dè gen kon m à propo d'la person kon m.j'sé kon chang pa 1 person mè on peu lui dir skon resen et ki sé ?
- Alors tu vien ou alors c est moi ki vien ! ?
- jpasré2m1 vR19H30ou av...ynpasrari1sita1cop.tve kjtemen manG ?tsé kla jssYpReureu.qd ons-parl,ttvatrobi1.Gldroi2tCduir ?T1amour.tpe etr fiRdetr lacoz2mn boner.jtM !
- OK j'arrive dans deux minutes mon bébé
- Bonne nuit a toi,celui qui sest emporté je pense que c est toi.si je me suis emporté c est pcq tu voulais pas mecouter.bisous jtm
- Ouvre moi la porte chui dans ta rue
- Bon allé derniere essai aps jabandonne,question banal :cmt ca va ?bonne soirée.

Corpus du DELIC

- Celui de derrière ou celui 2 l'entrée principale...
- Kikoo viens d'avoir ton msg... 2min chui libre entre 12 et 14h... Bonne soirée bisous à 2min...
- Kikoo té pas mort j'espère... Té o courant ki a 1 soirée au <NOM> ce week moi j'y serai po mais si t'y vas passe le bonjour o potes... A bientôt biz mràou...
- Kikoo tu devineras jamais avec ki g mangé à midi... Suspense... Roulement de tambour... <NOM>...Elle te fais des Bisous... Et moi aussi à bientôt grou...
- Kikoo à quelle heure tu finis ? Pcq je serai déjà sur manosque... Biz grou...
- bjr, bjr, personne pour m'amener o train... Ma mère partait à 8h... J'espère ke le cours sera pas tro cho... Si té pas tro occupé et si tu veux bien on fera 1 sceance 2 rattrapage... Bonne journée bisous grou...
- Kikoo i'a pas 2 soucis profite bien 2 ton chéri. Ce week jvé voir <NOM>. Des bisous à bientôt grou...
- Kikoo je viens 2 partir... J'arrive vers 11h15... Préviens moi si té pas chez toi... Et on se rejoint à la fac. Sinon je monte direct chez toi bisous à toute... grou...
- Mici bisous à 2main graou...
- Kikoo viens d'avoir ton msg g appelé lio... Moi ossi g hate bisous grou...
- oki, kool...
- Kikoo viens d'avoir ton msg... Je rentre aujourd'hui... Désolé bisous grou...
- J'espère qui te reste plus bcp 2 route... Je dormirai p-e qd tu seras rentré... Bisous
- Kikoo g pas u le temps 2 retirer... faudra y aller qd j'arrive... Désolé...
- G 25min 2 retard donc vers 18h55
- C des sous qd mm... mais on va trouver...
- COUCOU JE SUIS A MANOSK CA SERA +SIMPLE PR S'APELé !TU DOIS ETRE A AIX COURAGE !BIZ
- Il te passe le bonjour aussi. Je vé po tarder. bisous à toute.
- coucou ti fréro j'esper k tu t pa tro ennuyé san nou... bon courage pr cet aprem et a ce soir. gros bibis
- envoi moi un ti mess pr me dire si t bien rentré merci d'etre paC ca ma fé plaisir bisous.

F

Expérimentation autour de l'indexation de rubriques métiers

F.1 Extraits du corpus de requêtes utilisées pour l'évaluation

boiron
formule1
formule 1
p&t
éveché Tours
académie Paris
draty
non-voyant
audiovisuelle
restaurant viet
ce
entérologue
gastrologue
salle de concert
bureau de tabac
Paris
Rennes
Tours
hôtel sans restaurant hôtel avec restaurant
stations-service
camping à la ferme
golf court
primeur
chasse
mourir
coiffure afro
téléphone portable
centre médico-sportif
école maternelle privée
danse africaine
salle de cinéma
salle de cinémas
employée de maison
foyer
livre-librairie
portable
vêtement fabrication
édition
agence de mode
administrateur de bien
administrateur de biens
clinique et hôpitaux
ramonage fumisterie
...

F.2 Sortie XML de TiLT pour l'indexation de la requête "sécurité sociale"

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<TILT>
    <IMOGENE>
        <PLEIN>cpam
        </PLEIN>
    </IMOGENE>
    <INDEXEUR>
        <RUBRIQUES HOMOGENEITE="1" INTERPRET="CAISSE PRIMAIRE ASSURANCE MALADIE" \\
        RANG="101"  REDONDANCE="0" SCORE="1">
        <RUBRIQUE ID="58353600">sécurité sociale
        </RUBRIQUE>
        <MULTIFORMULE EVALUATION="correct">
            <THEME bruts="cpam" formes="cpam" lemmes="CPAM" libelle="CAISSE PRIMAIRE ASSURANCE\\
             MALADIE" nom="O_CAISSE_PRIMAIRE_D_ASSURANCE_MALADIE_1" rangs_elementaires=\\
            "101"/>
                <LISTE_RUBRIQUES>58353600
                </LISTE_RUBRIQUES>
                <CRITERES_BESLISSING>
                    <CRITERE identifiant="scoreMultiformule" type="Numerique" valeur="1"/>
                    <CRITERE identifiant="redondanceMultiformule" type="Numerique" valeur="0"/>
                    <CRITERE identifiant="homogeneiteMultiformule" type="Numerique" valeur="1"/>
                    <CRITERE identifiant="rangMultiformule" type="Numerique" valeur="101"/>
                </CRITERES_BESLISSING>
            </MULTIFORMULE>
            <MULTIFORMULE EVALUATION="correct">
                <THEME bruts="cpam" formes="cpam" lemmes="CPAM" libelle="CAISSE PRIMAIRE ASSURANCE\\
                MALADIE" nom="O_CAISSE_PRIMAIRE_D_ASSURANCE_MALADIE_1" rangs_elementaires="101"/>
                <LISTE_RUBRIQUES>58353600
                </LISTE_RUBRIQUES>
                <CRITERES_BESLISSING>
                    <CRITERE identifiant="scoreMultiformule" type="Numerique" valeur="1"/>
                    <CRITERE identifiant="redondanceMultiformule" type="Numerique" valeur="0"/>
                    <CRITERE identifiant="homogeneiteMultiformule" type="Numerique" valeur="1"/>
                    <CRITERE identifiant="rangMultiformule" type="Numerique" valeur="101"/>
                </CRITERES_BESLISSING>
            </MULTIFORMULE>
            <MULTIFORMULE EVALUATION="indifferent">
                <THEME bruts="cpam" formes="cpam" lemmes="CPAM" libelle="SECURITE SOCIALE" \\
                nom="O_SECURITE_SOCIAL_1" rangs_elementaires="104"/>
                <LISTE_RUBRIQUES>58353600
                </LISTE_RUBRIQUES>
                <CRITERES_BESLISSING>
                    <CRITERE identifiant="scoreMultiformule" type="Numerique" valeur="1"/>
                    <CRITERE identifiant="redondanceMultiformule" type="Numerique" valeur="0"/>
                    <CRITERE identifiant="homogeneiteMultiformule" type="Numerique" valeur="1"/>
                    <CRITERE identifiant="rangMultiformule" type="Numerique" valeur="104"/>
                </CRITERES_BESLISSING>
            </MULTIFORMULE>
</CLASSEMENT>
</TRACES>
```

```
</INDEXEUR>
</TILT>
```

Bibliographie

[Abney, 1989] S.P. Abney. A computational Model of Human Parsing. *Journal of psycholinguistic Research*, 1989.

[Abney, 1991] S. Abney. *Parsing by chunks.* Kluwer Academic, 1991.

[Agirre and Rigau, 1996] E. Agirre and G. Rigau. Word sense disambiguation using Conceptual Density. In *Proceedings of the 16th conference on Computational linguistics*, 1996.

[Akiba *et al.*, 2002] Y. Akiba, T. Watanabe, and E. Sumita. Using Language and Translation Models to Select the Best among Outputs from Multiple MT Systems. In *Proceedings of the 19th international conference on Computational linguistics*, 2002.

[Altmann, 1998] G. T. M. Altmann. Ambiguity in Sentence Processing. *Trends in Cognitive Sciences*, 2(4) :146–152, April 1998.

[Audibert, 2003] L. Audibert. Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences. In *Actes de la conférence TALN*, 2003.

[Audibert, 2007] L. Audibert. Désambiguïsation lexicale automatique : sélection automatique d'indices. In *Actes de la conférence TALN*, 2007.

[Bachimont, 1992] B. Bachimont. *Le contrôle dans les systèmes à base de connaissances.* HERMES, 1992.

[Benabbou *et al.*, 2004] L. Benabbou, A. Guitouni, and N. Belacel. Algorithme d'apprentissage pour inférer les paramètres de PROAFNT. *ASAC*, 2004.

[Bilhaut, 2003] F. Bilhaut. The LinguaStream platform. In *Proceedings of Spanish Society for Natural Language Processing Conference*, 2003.

[Biskri and Delisle, 1999] I. Biskri and S. Delisle. Un modèle hybride pour le textual data mining : un mariage de raison entre le numérique et le linguistique. In *Actes de la conférence TALN*, 1999.

[Biskri *et al.*, 1997] I. Biskri, C. Jouis, F. Le Priol, J.P. Descles, J.G. Meunier, and W. Mustafa. Outil d'aide à la fouille documentaire : approche hybride numérique linguistique. In *Actes de la conférence internationale : Linguistique et Informatique : Théories et Outils pour le traitement Automatique des Langues*, 1997.

[Blache and Rauzy, 2006] P. Blache and S. Rauzy. Mécanismes de contrôle pour l'analyse en grammaires de propriétés. In *Actes de la conférence TALN*, 2006.

[Blache, 2000] P. Blache. Le rôle des contraintes dans les théories linguistiques et leur intérêt pour l'analyse automatique : les Grammaires de Propriétés. In *Actes de la conférence TALN*, 2000.

[Bourigault and Frérot, 2004] D. Bourigault and C. Frérot. Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène. In *Actes de la conférence TALN*, 2004.

[Bourigault *et al.*, 2005] D. Bourigault, C. Fabre, C. Frérot, M.P. Jacques, and S. Ozdowska. Syntex, analyseur syntaxique de corpus. In *Actes de la conférence TALN*, 2005.

[Bourigault, 2007] D. Bourigault. *Un analyseur syntaxique opérationnel : SYNTEX.* PhD thesis, Université Toulouse Le Mirail, 2007.

[Boutilier *et al.*, 2004] C. Boutilier, R. Brafman, C. Domshlak, H.H. Hoos, and D. Poole. CP-nets : A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research*, 2004.

[Bouyssou and Vincke, 2006] D. Bouyssou and P. Vincke. Relations binaires et modélisation des préférences. *Concepts et Méthodes pour l'Aide à la Décision*, Chap. 1, 2006.

[Bouyssou, 2001] D. Bouyssou. Outranking approach. *Encyclopedia of optimization*, 2001.

[Bové *et al.*, 2006] R. Bové, C. Chardenon, and J. Véronis. Prise en compte des disfluences dans un système d'analyse syntaxique automatique de l'oral. In *Actes de la conférence TALN*, 2006.

[Bové, 2005] R. Bové. Étude de quelques problèmes de phonétisation dans un système de synthèse de la parole à partir de SMS. In *Actes de la conférence TALN*, 2005.

[Brans and Vincke, 1985] J.P. Brans and P. Vincke. A preference ranking organization method. *Management Science*, 35(6) :647–656, 1985.

[Briscoe and Carroll, 1993] T. Briscoe and J. Carroll. Generalized probabilistic LR parsing of natural language (Corpora) with unification-based grammars. In *Computational Linguistics*, 1993.

[Callison-Burch and Flournoy, 2001] C. Callison-Burch and R. S. Flournoy. A program for automatically selecting the best output from multiple machine translation engines. In *Proceedings of MT Summit VIII*, 2001.

[Cao and Li, 2002] Y. Cao and H. Li. Base Noun Phrase translation using web data and the EM algorithm. In *Proceedings of the 19th international conference on Computational linguistics*, 2002.

[Cardey-Greenfield, 1996] S. Cardey-Greenfield. *L'Ambiguïté*. BULAG, 1996.

[Charniak, 2000] E. Charniak. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL*, 2000.

[Charniak, 2005] E. Charniak. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the ACL*, 2005.

[Chaudhuri and Dayal, 1997] S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec.*, 1997.

[Chieu and Ng, 2002] H.L. Chieu and H.T. Ng. A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text. In *Eighteenth national conference on Artificial intelligence*, 2002.

[Chinchor and Hirschmann, 1997] N. Chinchor and L. Hirschmann. MUC-7 Coreference Task definition, Version 3.0. In *Proceedings of MUC-7*, 1997.

[Church, 1982] K. Church. Coping with Syntactic Ambiguity or How to Put the Block in the Box on the Table. *American Journal of Computational Linguistics*, 1982.

[Chvatal, 1973] V. Chvatal. On the computational complexity of finding a kernel. Centre de recherches mathematiques, 1973.

[Claveau, 2003] V. Claveau. Extraction de couples nom-verbe sémantiquement liés : une technique symbolique automatique. In *Actes de la conférence TALN*, 2003.

[Collins and Koo, 2005] M. Collins and T. Koo. Discriminative reranking for natural language parsing. *Computational Linguistics*, 2005.

[Collins, 1997] M. Collins. Three generative, lexicalised models for statistical parsing. In *The Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997.

[Collins, 2003] M. Collins. Head-Driven Statistical Models for Natural Language Parsing. *Association for Computational Linguistics*, 2003.

[Copestake and Lascarides, 1997] A. Copestake and A. Lascarides. Integrating Symbolic and Statistical Representations : The Lexicon Pragmatics Interface. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997.

[Cunningham *et al.*, 2001] H. Cunningham, D. Maynard, V. Tablan, C. Ursu, and K. Bontcheva. Developing Language Processing Components with GATE, 2001.

[Dang and Palmer, 2002] H.T. Dang and M. Palmer. Combining Contextual Features for Word Sens Disambiguation. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation*, 2002.

[Dias and Mousseau, 2006] L.C. Dias and V. Mousseau. Inferring Electre's veto parameters from outranking examples. *European Journal of Operational Research*, 2006.

[Dubois *et al.*, 2004] D. Dubois, S. Kaci, and H. Prade. CP-nets and possibilistic logic : Two approaches to preference modeling. Steps towards a comparison. *Computational Intelligence*, 2004.

[Erman *et al.*, 1980] L.D. Erman, F. Hayes-Roth, V.R. Lesser, and D.R. Reddy. The Hearsay-II Speech-Understanding System : Integrating Knowledge to Resolve Uncertainty. *Computing Surveys*, 12(2), June 1980.

[et Alexandre Patry, 2007] Philippe Langlais et Alexandre Patry. Enrichissement d'un lexique bilingue par analogie. In *Actes de la conférence TALN*, 2007.

[Fairon and Paumier, 2006] C. Fairon and S. Paumier. A translated corpus of 30,000 French SMS. In *Proceedings of LREC2006*, 2006.

[Farah and Vanderpooten, 2006] M. Farah and D. Vanderpooten. A Multiple Criteria Approach for Information Retrieval. *String Processing and Information Retrieval*, 2006.

[Forney, 1973] G.D. Forney. The Viterbi algorithm. In *Proceedings of the IEEE*, volume 61, 1973.

[Frérot *et al.*, 2003] C. Frérot, D. Bourigault, and C. Fabre. Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. *Traitement automatique des langues*, 2003.

[Gamma *et al.*, 1980] E. Gamma, R. Johnson, and R. Helm. *Design patterns : Elements of reusable object-oriented software*. Addison-Wesley, 1980.

[Gibson and Pearlmutter, 1998] E. Gibson and N.J. Pearlmutter. Constraints on sentence comprehension. *Trends in Cognitive Sciences*, 1998.

[Grabar and Zweigenbaum, 2005] N. Grabar and P. Zweigenbaum. Utilisation de corpus de spécialité pour le filtrage de synonymes de la langue générale. In *Actes de la conférence TALN*, 2005.

[Grabisch and Perny, 2002] M. Grabisch and P. Perny. Agrégation Multicritère. *Logique floue, principes, aide à la décision*, 2002.

[Greco *et al.*, 2001] S. Greco, B. Matarazzo, and R. Slowinski. Rough sets for multicriteria decision analysis. *European Journal of Operational Research*, 2001.

[Guimier De Neef and Fessard, 2007] E. Guimier De Neef and S. Fessard. Évaluation d'un système de transcription de SMS. In *26th conference on Lexis and Grammar*, 2007.

[Guimier De Neef *et al.*, 2002] E. Guimier De Neef, M. Boualem, C. Chardenon, P. Filoche, and J. Vinesse. Natural language processing software tools and linguistic data developed by france télécom r&d. In *Indo European Conference on Multilingual Technologies (IECMT)*, 2002.

[Guimier De Neef *et al.*, 2007] E. Guimier De Neef, A., and J. Park. TiLT correcteur de SMS : évaluation et bilan qualitatif. In *Actes de la conférence TALN*, 2007.

[Guitouni and Martel, 1998] A. Guitouni and J.M. Martel. Tentative guidelines to help choosing an appropriate MCDA method. *European Journal of Operational Research*, 1998.

[Hansson, 1994] S.O. Hansson. Decision Theory : A Brief Introduction, 1994.

[Hocq, 2006] S. Hocq. Étude des SMS en français : constitution et exploitation d'un corpus aligné SMS - langue standard. Technical report, Aix-en-Provence, rapport de Master II Industries des langues, 2006.

[Ide and Véronis, 1998] N. Ide and J. Véronis. Introduction to the special issue on word sense disambiguation : the state of the art. In *Computational Linguistics*, 1998.

[Kononenko, 1994] I. Kononenko. Estimating attributes : Analysis and extensions of RELIEF. In *Proceedings of the European Conference on Machine Learning*, 1994.

[Lebarbé, 2001] T. Lebarbé. Vers une plate-forme multi-agents pour l'exploration et le traitement linguistiques. In *Actes de la conférence TALN*, 2001.

[Leclercq, 1984] J.P. Leclercq. Propositions d'extension de la notion de dominance en présence de relations d'ordre sur les pseudo-critères : MELCHIOR. *Revue Belge de Recherche Opérationnelle, de Statistique et d'Informatique*, 24 :32–46, 1984.

[Lopez *et al.*, 2002] P. Lopez, C. Fay-Varnier, and A. Roussanaly. Lexicalized Grammar Specialization for Restricted Applicative Languages. In *Actes de la conférence LREC*, 2002.

[Martin and Legret, 2005] C. Martin and M. Legret. La méthode multicritère ELECTRE III : Définitions, principe et exemple d'application à la gestion des eaux pluviales en milieu urbain. *Bulletin des laboratoires des Ponts et Chaussées*, 2005.

[McCarthy and Lehnert, 1995] J.F. McCarthy and W.G. Lehnert. Using Decision Trees for Coreference Resolution. In *Proceedings of the fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, 1995.

[McCarthy *et al.*, 2004] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.

[McRoy, 1992] S.W. McRoy. Using Multiple Knowledge Sources for Word Sense Discrimination. *Computational Linguistics*, 1992.

[Mierswa *et al.*, 2003] I. Mierswa, R. Klinkenberg, S. Fischer, and O. Ritthoff. AFlexible Platform for Knowledge Discovery Experiments : YALE – Yet Another Learning Environment. Technical report, Collaborative Research Center on Computational Intelligence, 2003.

[Miller, 1995] G.A. Miller. WordNet : a lexical database for English. *Commun. ACM*, 1995.

[Mintzberg *et al.*, 1976] H. Mintzberg, D. Raisinghani, and A. Théorêt. The Structure of 'Unstructured' Decision Processes. *Administrative Sciences Quaterly*, 1976.

[Mitkov, 1998] R. Mitkov. Robust pronoun resolution with limited knwoledge. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, 1998.

[Mousseau and Slowinski, 1998] V. Mousseau and R. Slowinski. Inferring an ELECTRE TRI model from assignment examples. *Journal of Global Optimization*, 1998.

[Mousseau *et al.*, 1999] V. Mousseau, R. Slowinski, and P. Zielniewitcz. Electre Tri 2.0 Methodological guide and user's manual. Technical report, LAMSADE, Paris Dauphine, 1999.

[Mousseau *et al.*, 2001] V. Mousseau, J. Figueira, and J.P. Naux. Using assignment examples to infer weights for ELECTRE TRI method : Some experimental results. In *European Journal of Operational Research*, 2001.

[Mousseau, 1995] V. Mousseau. Eliciting information concerning the relative importance of criteria. In *Advances in Multicriteria Analysis*. Kluwer Academic, 1995.

[Mousseau, 2003] V. Mousseau. *Élicitation des préférences pour l'aide multicritère à la décision*. PhD thesis, Université Paris Dauphine, 2003.

[Nastase and Szpakowitcz, 2001] V. Nastase and S. Szpakowitcz. Word Sense Disambiguation in Roget's Thesaurus. *Workshop on WordNet and Other Lexical Resources*, 2001.

[Ngo The and Mousseau, 2002] A. Ngo The and V. Mousseau. Using assignment examples to infer category limits for the ELECTRE TRI method. *Journal of Multicriteria Decision Analysis*, 2002.

[Niv, 1994] M. Niv. A psycholinguistically motivated parser for CCG. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994.

[Paiva and Evans, 2005] D.S. Paiva and R. Evans. Empirically-based Control of Natural Language Generation. In *Proceedings of the 43rd Annual Meeting of the ACL*, 2005.

[Paroubek *et al.*, 2005] P. Paroubek, L.-G. Pouillot, I. Robba, and A. Vilnat. EASY : Campagne d'évaluation des analyseurs syntaxiques. In *Proceedings of the 12ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*, 2005.

[Péry-Woodley, 1995] M.P. Péry-Woodley. Quels corpus pour quels traitements automatiques ? *Traitement automatique des langues*, 1995.

[Piron, 1994] C. Piron. *Le défi des langues : Du gâchis au bon sens.* L'Harmattan, 1994.

[Pusateri and Thong, 2001] E. Pusateri and J.M. Van Thong. N-best List Generation using Word and Phoneme Recognition Fusion. In *Proceedings of the European Conference on Speech*, 2001.

[Rady, 1983] M. Rady. *L'ambiguïté des langues est-elle à l'origine de l'indéterminisme des procédures de traitement ?* PhD thesis, LIMSI, 1983.

[Ratnaparkhi, 1996] A. Ratnaparkhi. A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1996.

[Rigau *et al.*, 1997] G. Rigau, J. Atserias, and E. Agirre. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997.

[Robertson and Walker, 2001] S. Robertson and S. Walker. Microsoft Cambridge at TREC-9 : Filtering track. In *The Ninth Text REtrieval Conference (TREC–9)*, 2001.

[Rosé and Waibel, 1994] C.P. Rosé and A. Waibel. Recovering From Parser Failures : A Hybrid Statistical/Symbolic Approach, 1994.

[Rosso *et al.*, 2003] P. Rosso, F. Masulli, and D. Buscaldi. Word sense disambiguation combining conceptual distance, frequency and gloss. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, 2003.

[Roy and Bouyssou, 1987] B. Roy and D. Bouyssou. Famille de critères : problème de cohérence et de dépendance. Documents du Lamsade num. 37, 1987.

[Roy and Bouyssou, 1993] B. Roy and D. Bouyssou. *Aide Multicritère à la Décision : Méthodes et Cas.* Economica, 1993.

[Roy and Mousseau, 1996] B. Roy and V. Mousseau. A theoritical framework for analysing the notion of relative importance of criteria. *Journal of Multicriteria Decision Analysis*, 1996.

[Roy, 1974] B. Roy. Critères multiples et modélisation des préférences : l'apport des relations de surclassement. *Revue d'économie politique*, 1974.

[Sabah, 1989] G. Sabah. *L'intelligence Artificielle et le Langage : processus de compréhension.* HERMES, 1989.

[Sabah, 1990] G. Sabah. CARAMEL : a flexible model for interaction between the cognitive processes underlying natural language understanding. *Proceedings of the 13th conference on Computational linguistics*, 1990.

[Sagot and de la Clergerie, 2006] B. Sagot and E. Villemonte de la Clergerie. Trouver le coupable : Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Actes de la conférence TALN*, 2006.

[Schubert, 1984] L.K. Schubert. On parsing preferences. In *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, 1984.

[Schwartz and Austin, 1991] R. Schwartz and S. Austin. A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses. In *ICASSP '91 : Proceedings of the Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 1991.

[Shen and Joshi, 2003] L. Shen and A.K. Joshi. An SVM based voting algorithm with application to parse reranking. In *Proceedings of CoNLL*, 2003.

[Shen *et al.*, 2004] L. Shen, A. Sarkar, and F.J. Och. Discriminative reranking for machine translation. In *Proceedings of the Joint HLT and NAACL Conference*, 2004.

[Shim *et al.*, 2002] J.P. Shim, M. Warkentin, J.F. Courtney, D.J. Power, R. Sharda, and C. Carlsson. Past, present, and future of decision support technology. *Journal of Decision Support Systems*, 2002.

[Skalka *et al.*, 1994] J.M. Skalka, D. Bouyssou, and D. Vallée. ELECTRE III-IV, version 3.x, Aspects Méthodologiques (Tome 1), Guide d'utilisation (Tome 2). Documents du lamsade n. 85 et 85bis, Université de Paris Dauphine, France, 1994.

[Stefanini and Demazeau, 1995] M.H. Stefanini and Y. Demazeau. TALISMAN : A Multi-Agent System for Natural Language Processing. In *SBIA '95 : Proceedings of the 12th Brazilian Symposium on Artificial Intelligence*, 1995.

[Tardif, 2005] O. Tardif. Annotation de la coréférence entre expressions référentielles. In *Actes de la quatrième journée de linguistique de corpus*, 2005.

[Tardif, 2006] O. Tardif. Résoudre la coréférence à l'aide d'un classifieur bayésien naïf. In *Actes de la conférence TALN*, 2006.

[Toussaint *et al.*, 1998] Y. Toussaint, F. Namer, B. Daille, C. Jacquemin, J. Royauté, and N. Hathout. Une approche linguistique et statistique pour l'analyse de l'information en corpus. In *Actes de la conférence TALN*, 1998.

[Tzoukermann *et al.*, 1995] E. Tzoukermann, D.R. Radev, and W.A. Gale. Combining Linguistic Knowledge and Statistical Learning in French Part-of-Speech Tagging. *EACLSIGDAT Workshop*, 1995.

[Vanderpooten, 1989] D. Vanderpooten. The interactive approach in MCDA : a technical framework and some basic conceptions. *Mathematical modelling*, 1989.

[Vergne, 2001] J. Vergne. Analyse syntaxique automatique de langues : du combinatoire au calculatoire. In *Actes de la conférence TALN*, 2001.

[Véronis and Guimier De Neef, 2006] J. Véronis and E. Guimier De Neef. Le traitement des nouvelles formes de communication écrite. *In Sabah, G. (Ed.), Compréhension automatique des langues et interaction*, 2006.

[Véronis, 2004] J. Véronis. Informatique et linguistique. Université de Provence, 2004.

[Vincke, 1989] P. Vincke. *L'Aide Multicritère à la Décision.* SMA, 1989.

[Weissenbacher and Nazarenko, 2007a] D. Weissenbacher and A. Nazarenko. Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. In *Actes de la conférence TALN*, 2007.

[Weissenbacher and Nazarenko, 2007b] D. Weissenbacher and N. Nazarenko. A bayesian approach combining surface clues and linguistic knowledge : Application to the anaphora resolution problem. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP'07)*, 2007.

[Whittemore and Ferrara, 1990] G. Whittemore and K. Ferrara. Empirical Study of Predictive Powers of Simple Attachment Schemes for Post-modifier Prepositional Phrases. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, 1990.

[Wilks, 1985] Y. Wilks. Right attachment and preference semantics. In *Proceedings of the second conference on European chapter of the Association for Computational Linguistics*, 1985.

[Wilks, 1998] Y. Wilks. Word Sense Disambiguation using Optimised Combinations of Knowledge Sources. In *Proceedings of the conference on Computational linguistics*, 1998.

[Yager, 1993] R.R. Yager. Families of OWA operators. *Fuzzy sets and systems*, 1993.

[Yarowsky, 1992] D. Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th conference on Computational linguistics*, 1992.

[Yarowsky, 1994] D. Yarowsky. Decision lists for lexical ambiguity resolution : application to accent restoration in Spanish and French. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994.

[Yates and Schoenmackers, 2006] A. Yates and S. Schoenmackers. Detecting Parser Errors Using Web-based Semantic Filters. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006.

[Yvon, 2007] F. Yvon. Une petite introduction au traitement automatique des langues. Telecom Paris, 2007.

Une approche par surclassement pour le contrôle d'un processus d'analyse linguistique

Les systèmes de Traitement Automatique des Langues Naturelles (TALN) sont de manière récurrente confrontés au problème de la génération et de la propagation d'hypothèses concurrentes et erronées. Afin d'écarter ces erreurs d'interprétation du processus d'analyse, il apparaît indispensable d'avoir recours à des stratégies spécifiques de contrôle dont l'objectif est de différencier les hypothèses concurrentes selon leur degré de pertinence. Sur la plupart des cas d'indétermination observés, on constate que cette évaluation de la pertinence relative des hypothèses repose sur l'exploitation de plusieurs sources de connaissances hétérogènes, qui doivent être combinées pour garantir un contrôle robuste et fiable. À partir de ce constat, nous avons montré que le traitement des indéterminations répondait à une formalisation générique en tant que problème décisionnel basé sur de multiples critères de comparaison. Cette formalisation et la recherche d'une méthodologie adaptée nous ont conduit vers une approche par surclassement issue des travaux en Aide MultiCritère à la Décision (AMCD). Par rapport aux méthodes alternatives, cette approche se différentie notamment par l'importance qu'elle accorde aux connaissances et préférences qu'un expert est en mesure d'apporter sur le problème traité.

À partir de cette intersection novatrice entre le TALN et l'AMCD, nos travaux se sont focalisés sur le développement d'un module décisionnel de contrôle multicritère. L'intégration de ce module au sein d'un système complet de TALN nous a permis d'attester d'une part la faisabilité de notre approche et d'autre part de l'expérimenter sur différents cas concrets d'indétermination.

Mots clés : LANGAGE NATUREL,TRAITEMENT DU (INFORMATIQUE), DÉCISION MULTICRITÈRE, AMBIGUÏTÉ, SYSTÈMES INFORMATIQUES - - ÉVALUATION

Controlling a linguistic analysis process using an outranking approach

Natural Language Processing (NLP) systems are continuously faced with the problem of generating concurrent hypotheses, of which some can be erroneous. In order to avoid the propagation of erroneous hypotheses, it appears to be essential to apply specific control strategies, which aim to distinguishing concurrent hypotheses based on their relevance.

On most of observed indetermination cases, we have noticed that multiple heterogeneous knowledge sources have to be combined to determine the hypotheses relative relevance. According to this observation, we show that the control of the indetermination cases can be formalised as a decisional process based on multiple criteria.

This decisional formalisation and our research of an adapted methodology have conducted us toward an outranking approach issued from the MultiCriteria Decision Aid (MCDA) paradigm. This approach differs from alternative methods by the importance granted to knowledge and preferences that an expert can express about a given problem.

From this innovative intersection between NLP and MCDA, our work has been focalised on the development of a decisional module dedicated to multicriteria control. The integration of this module into a complete NLP system has allowed us to attest the feasibility of our approach and to perform experimentation on concrete indetermination cases.

Key words : NATURAL LANGUAGE, PROCESSING (COMPUTING), MULTICRITERIA DECISION, COMPUTING SYSTEMS – EVALUATION

Discipline : Informatique

Laboratoire : Université de Caen
G.R.E.Y.C. équipe ISLanD
Campus Côte de Nacre
boulevard du Maréchal Juin
BP 5186 - 14032 Caen CEDEX

France Télécom division R&D
TECH/EASY/LN
2, av. Pierre Marzin
F-22307 Lannion cedex