



LIRMM
UMR 5506
161, rue Ada
34394 Montpellier Cedex 5

FRANÇOIS
DUPRESSOIR
ÉNSL
2005-2006

RAPPORT DE STAGE

Exploitation de la construction syntaxique des verbes pour l'évaluation automatique de l'influence sémantique de leurs compléments.



LIRMM
UMR 5506
161, rue Ada
34394 Montpellier Cedex 5

FRANÇOIS
DUPRESSOIR
ÉNSL
2005-2006

RÉSUMÉ :

Une tâche importante dans le cadre d'un travail de contraction de phrase, ou de résumé automatique en général, est le repérage des éléments syntaxiques obligatoires, mais aussi des éléments sémantiques importants.

Après une courte présentation du laboratoire, de l'équipe, et des logiciels utilisés par celle-ci, je présenterai dans un premier temps les concepts grammaticaux et linguistiques utiles à la compréhension et à la résolution du problème, puis la solution envisagée pour le résoudre, en commençant par la constitution d'une ressource lexicale complète et en terminant par une utilisation possible d'une telle ressource dans le cadre du laboratoire par la constitution d'une grammaire.

MOTS-CLEFS :

contraction de phrases, traitement automatique des langues, langues naturelles, circonstants, sous-catégorisation, grammaires d'unification

Remerciements

Merci à Mehdi YOUSFI-MONOD, Violaine PRINCE et Jacques CHAUCHÉ pour leur soutien en ce qui concerne l'utilisation de SYGMART et SYGFRAN et pour la patience et la compréhension dont ils ont fait preuve lorsque je me suis retrouvé face aux problèmes liés à la complexité de la langue. Un grand merci à Augusta MELA pour ses lumières et ses références autant en grammaire qu'en linguistique, et en particulier en TALN.

Merci à Caroline DAVID, qui m'a confirmé, si besoin était, que la linguistique était une discipline formidable mais "un peu compliquée".

Merci à Virginie QUESNAY pour la classe `rapportiup`, utilisée pour mettre en forme ce rapport.

Merci à Judicaëlle pour le reste, et pour ça aussi.

Ce rapport, produit avec L^AT_EX 2_ε, a été compilé le 30 août 2006.

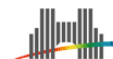
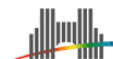


Table des matières

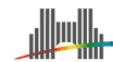
Remerciements	i
Liste des illustrations	v
Présentation	1
1 Présentation du LIRMM	1
1.1 Présentation Générale	1
1.2 Le département Informatique	1
2 Présentation de l'équipe de Traitement Automatique des Langues	2
1 Contexte Scientifique	5
1.1 Analyse morpho-syntaxique	5
1.1.1 Le système SYGMART	5
1.1.2 Les règles SYGFRAN	6
1.1.3 Résultats	6
1.2 Résumé automatique	8
1.2.1 L'extraction de phrases	8
1.2.2 L'extraction de constituants	8
1.3 Contraction de phrase	9
1.3.1 Présentation du système	9
1.3.2 Résultats	10
Problématique	11
2 Considérations Grammaticales	13
2.1 Grammaires traditionnelles	13
2.2 Grammaires d'unification	14
2.2.1 <i>Lexical functional grammar</i> (LFG)	14
2.2.2 Le Lexique-Grammaire	15
3 Marquage des compléments	19
3.1 Principes et choix théoriques	19
3.1.1 Choix d'une approche particulière	19
3.1.2 Constitution du lexique	19
3.2 Réalisation du marquage	22
3.2.1 Pré-traitement : Mise en forme de l'analyse SYGFRAN	23
Formes réfléchies	23
Formes impersonnelles	23
Voix passive	24
3.2.2 Repérage des compléments potentiels	25
3.2.3 Traitement : lecture du lexique et marquage des compléments	25



Détection des formes supports	26
Marquage effectif	26
3.3 Résultats et développements futurs	27
Conclusion	29
Annexes	31
A Extrait d'un conte polynésien	33
A.1 Texte Initial	33
A.2 Résultat de la compression	35
Bibliographie	37

Liste des illustrations

1.1	SYGFRAN fonctionne.	6
1.2	Analyse SYGFRAN partielle	7
1.3	Analyse d'une phrase ambiguë	7
2.1	Structures c	15
2.2	Structures f	15
2.3	Exemple d'unification	16
3.1	Extrait du lexique	21
3.2	Règle : propagation des formes impersonnelles infinitives	24
3.3	Règle : traitement des verbes supports	26



Présentation

1 Présentation du LIRMM

1.1 Présentation Générale

Le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier - LIRMM - est une unité mixte de recherche, dépendant conjointement de l'Université Montpellier II et du Centre National de la Recherche Scientifique. Les recherches actuelles et en émergence au LIRMM couvrent un large spectre de l'informatique et de ses applications :

- l'informatique fondamentale,
- l'interaction entre les systèmes informatiques et les utilisateurs,
- le développement de machines communicantes d'intervention, de production ou de service,
- le développement des composants matériels et logiciels des systèmes informatiques et de communication.

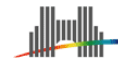
Treize ans après sa création en 1992, les interactions entre chercheurs de cultures initiales différentes ont conduit à de nouveaux thèmes de recherche dans lesquels les aspects logiciels et matériels sont abordés conjointement. Les recherches du LIRMM trouvent généralement une finalisation dans des domaines applicatifs aussi divers que la biologie, la chimie, les télécommunications, le secteur médical, la documentation...et dans les domaines propres du laboratoire : l'informatique, l'électronique et l'automatique. Le laboratoire regroupe 292 personnes (dont 152 permanents) :

- 95 enseignants-chercheurs et 29 chercheurs CNRS (et également INRIA,...)
- 28 ingénieurs, techniciens ou administratifs (+6 contractuels)
- 134 doctorants
- 3 chercheurs contractuels (hors doctorants).

Le LIRMM pilote deux formations doctorales, l'une en Informatique, l'autre en Systèmes automatiques et microélectroniques. En moyenne, la production scientifique annuelle du LIRMM est de 350 publications, dont 25 thèses de doctorat et 170 publications dans des revues ou des congrès d'audience internationale. Aux soutiens du CNRS et de l'Université Montpellier II, s'ajoutent ceux d'une quinzaine de programmes de recherche nationaux et d'une dizaine de programmes de recherche européens auxquels le LIRMM participe.

1.2 Le département Informatique

Le Département d'Informatique du LIRMM regroupe actuellement 76 chercheurs permanents, 13 chercheurs associés et plus de 70 doctorants, et quelques ingénieurs contractuels supportent les équipes dans le cadre de contrats européens ou industriels, ou d'actions nationales. Les thématiques du département couvrent l'essentiel de la recherche actuelle en informatique et ses applications :



- **Algorithmique** :
bioinformatique, cryptographie, graphes, réseaux,
- **Bases de Données et Systèmes d'Information** :
intégration de données, fouille de données, maintien de la cohérence,
- **Génie Logiciel** :
langages de programmation, objets, composants, modèles,
- **Intelligence Artificielle** :
apprentissage, contraintes, représentation des connaissances, systèmes multi-agents,
- **Interaction Homme-Machine** :
hypermedia, langage naturel, visualisation, web sémantique et e-learning.

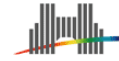
2 Présentation de l'équipe de Traitement Automatique des Langues

C'est dans ce contexte très riche que travaille l'équipe TAL, dont l'objectif est de concevoir et de réaliser un artefact qui soit en mesure d'accepter des productions langagières, et de les interpréter en vue de la réalisation de tâches précises : traduction, recherche d'information, classification de documents, dialogues, commandes de robots, etc. L'équipe se compose de :

- Jacques CHAUCHÉ (chercheur)
- Mathieu LAFOURCADE (chercheur)
- Violaine PRINCE (chercheuse)
- Mathieu ROCHE (chercheur)
- Anne PRELLER (chercheuse)
- Alain JOUBERT (chercheur)
- Sylvain DEGEILH (doctorant)
- Mehdi YOUSFI-MONOD (doctorant)
- Alexandre LABADIÉ (doctorant)

L'équipe travaille essentiellement sur le Français, mais avec des incursions dans d'autres langues (Anglais, Allemand), essentiellement dans le cadre de la traduction automatique. Le thème de recherche peut se décomposer en trois axes principaux :

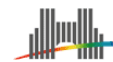
- **La Syntaxe**
qui se définit à travers deux impératifs :
 - Appréhension : définir le passage d'une structure S1 dans un modèle M1 vers une structure S2 d'un modèle M2. Par exemple, le passage d'une production en langue naturelle vers une expression logique.
 - Compréhension : définir, à partir de toutes les structures d'un modèle M1, les structures correspondantes qui doivent être obtenues par appréhension. Il s'agit d'une application.
- **La Traduction** ou Transformation, qui définit le passage d'une structure S1 dans un modèle M vers une structure S2 du même modèle M. Par exemple, la transformation d'une structure syntaxique arborescente d'une phrase dans une langue L1 vers une structure syntaxique arborescente dans une langue L2.
- **La Sémantique** ou Corrélation : Les structures d'un modèle M1 sont plongées dans un modèle M2 tel que l'on puisse définir des normes et des distances. Des calculs de "proximité" de structures permettent de traduire les relations sémantiques de la linguistique.



Les recherches sur ces trois axes donnent lieu à diverses applications :

- recherche d'informations à l'aide du langage naturel (CHAUCHÉ, LAFOURCADE, PRINCE, ROCHE, LABADIÉ)
- classification de documents par l'analyse de contenu (CHAUCHÉ, PRINCE, ROCHE)
- segmentation thématique de textes (CHAUCHÉ, PRINCE, ROCHE, LABADIÉ)
- création et amélioration de ressources lexicales (LAFOURCADE, PRINCE)
- amélioration de ressources multilingues (LAFOURCADE)
- traduction automatique fondée sur l'analyse (CHAUCHÉ, PRINCE)
- vérification grammaticale (PRELLER, PRINCE, DEGEILH)
- contraction automatique de textes (PRINCE, YOUSFI-MONOD)

C'est plus particulièrement sur cette dernière application que j'ai travaillé, sous la direction de Mehdi YOUSFI-MONOD et Augusta MELA, chercheuse associée à l'équipe. Mais avant de présenter mon travail, voici tout d'abord une présentation plus précise du contexte scientifique dans lequel s'inscrit le travail effectué lors du stage.



Chapitre 1

Contexte Scientifique

1.1 Analyse morpho-syntaxique

L'analyse morpho-syntaxique consiste à donner sur les éléments d'un texte des informations morphologiques (temps, genre, nombre...) et syntaxiques (nature, fonction...). Cette analyse est la base de toute application en traitement automatique des langues, naturelles ou non. Il existe plusieurs moyens de l'effectuer, mais je n'entrerai pas dans les détails, mon travail ne porte pas sur cette partie du traitement, mais un lien étroit existe entre l'analyse syntaxique et l'analyse sémantique. Je présenterai donc sommairement le système SYGMART, ainsi que l'ensemble de règles SYGFRAN utilisés au cours du stage sans passer en revue les différentes approches existantes pour l'analyse.

1.1.1 Le système SYGMART

SYGMART, introduit pour la première fois par Jacques CHAUCHÉ ([Cha84]), est un système de transformation d'éléments structurés qui peut servir à diverses opérations sur les chaînes de caractères, entre autres leur analyse syntaxique. Il se présente sous la forme de trois sous-systèmes :

- OPALE réalise le passage entre les chaînes de caractères et les éléments structurés manipulés par le système ;
- TELESY réalise les manipulations d'éléments structurés ;
- AGATE réalise le passage d'un élément structuré au format de sortie souhaité pour le système (chaîne de caractère, fichier texte, fichier XML...)

Les "éléments structurés" manipulés par SYGMART sont des arbres multi-étiquetés. Pour effectuer les transformations, SYGMART utilise, après les avoir compilés, un ensemble de règles de grammaire et des dictionnaires, fournis par l'utilisateur, et écrits dans le langage spécifique au sous-système qui les utilisera. Ces langages sont décrits dans le manuel de référence ([Cha01]), qui explique aussi plus finement le fonctionnement même des transformations, sans pour autant spécifier leur implémentation, dont il est plus particulièrement question dans [Cha84].

Une telle architecture en couches a plusieurs avantages, d'abord en permettant à l'utilisateur d'effectuer toutes les opérations qu'il souhaite effectuer avant la sortie finale en faisant simplement appel à TELESY autant de fois que nécessaires avec les grammaires de son choix, mais aussi en permettant de modifier ou d'ajouter une grammaire (et donc le traitement qui lui correspond) sans modifier les autres.

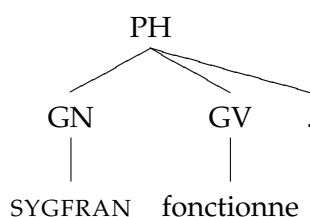
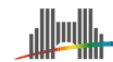


FIG. 1.1 – Un exemple d’analyse SYGFRAN réussie.

1.1.2 Les règles SYGFRAN

Pour effectuer l’analyse morpho-syntaxique d’un texte écrit en français, Jacques CHAUCHÉ, a écrit et met régulièrement à jour un ensemble de grammaires et de dictionnaires pour SYGMART : SYGFRAN. Ces règles et ces lexiques sont complétés au fil des expériences sur différents corpus et en fonction des demandes des utilisateurs, si bien que SYGFRAN compte actuellement plus de 12000 règles de grammaire et présente des résultats bien supérieurs¹ à d’autres systèmes d’analyse, même s’il est souvent difficile de comparer des résultats dans un domaine aussi ambigu et aussi peu formalisé que la grammaire d’une langue naturelle. Le principal inconvénient d’une telle méthode de construction de la ressource lexicale est qu’elle consiste en une juxtaposition de cas particuliers. Comme nous le verrons, intégrer des règles générales utilisant les récentes avancées en matière de grammaire et de linguistique dans un tel cadre est difficile. Cependant, il faut garder à l’esprit que SYGMART n’était, à l’origine qu’un système expérimental visant à démontrer l’efficacité de l’approche choisie pour l’analyse et qu’il a, à ce titre, fait ses preuves.

1.1.3 Résultats

SYGFRAN peut renvoyer les résultats sous diverses formes suivant la grammaire AGATE choisie. Voici une présentation succincte de quelques possibilités de SYGFRAN sous forme d’arbre syntaxique. Les étiquettes représentées ici sont incomplètes et les éléments réellement manipulés par les système portent bien plus d’informations, au cours du traitement comme dans le résultat final.

Lorsque SYGFRAN fonctionne, il renvoie le résultat sous la forme choisie². La figure 1.1 montre un exemple (très) simple d’analyse réussie.

En cas d’échec de l’analyse, SYGFRAN renvoie une analyse partielle (figure 1.2) qui peut être utilisée dans une application.

Enfin, lorsqu’il n’est pas possible de lever totalement les ambiguïtés de l’analyse, SYGFRAN dédouble l’arbre (figure 1.3) pour rendre compte des différentes interprétations possibles de la phrase, laissant l’utilisateur faire le choix, soit de manière manuelle, soit en écrivant des règles tenant compte du contexte, du thème général, du type de texte ou d’autres facteurs permettant une désambiguïsation.

Ces deux dernières caractéristiques de SYGFRAN (figures 1.2 et 1.3) font de lui un

¹Les résultats de la campagne d’évaluation EASY des analyseurs du français sont significatifs, SYGFRAN donnant des réponses justes à 70% pour les analyses en constituants contre 30% en moyenne pour les autres évaluateurs.

²Suivant la grammaire AGATE sélectionnée.

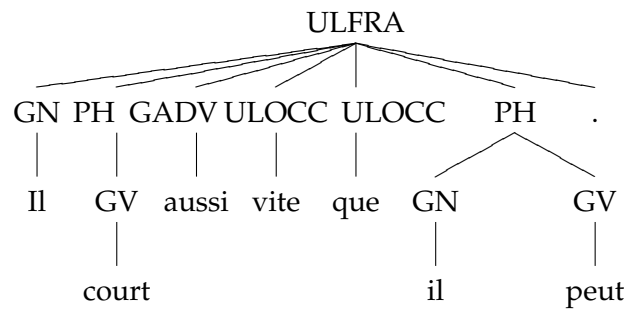
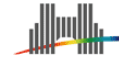


FIG. 1.2 – Un exemple d’analyse partielle : *Il court aussi vite qu’il peut.*

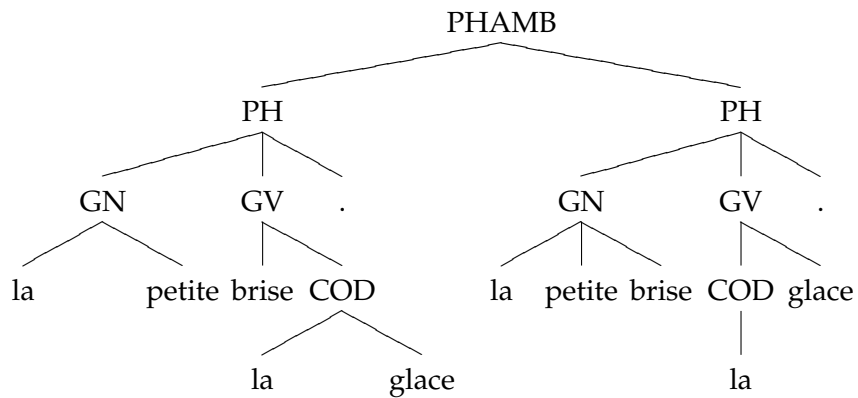
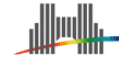


FIG. 1.3 – Un exemple d’analyse ambiguë : *La petite brise la glace.*



analyseur syntaxique robuste, qui peut servir à des applications alors même que l'analyse qu'il fait de la phrase n'est pas complète ou reste ambiguë.

C'est donc sur ce système robuste et performant que sont effectués la plupart des travaux de l'équipe TAL du LIRMM, et en particulier le travail de résumé automatique. Je vais tout d'abord présenter le travail de résumé de manière plus générale avant de présenter plus particulièrement l'approche choisie dans l'équipe.

1.2 Résumé automatique

Parmi toutes les tâches du traitement automatique des langues, le résumé automatique est l'une des plus compliquées. En effet, elle fait intervenir des considérations aussi bien syntaxiques et lexicales que sémantiques et, par sa nature de réduction d'information, pose de gros problèmes conceptuels quant à la mesure de "l'importance" d'un élément ou de sa nécessité. De plus, ses applications en font un noeud important parmi les nombreux problèmes liés au TAL.

Plusieurs approches sont envisagées pour résumer un texte de manière automatique. Toutes possèdent leurs avantages et leurs inconvénients, et la meilleure des solutions serait sans aucun doute de les appliquer toutes ou presque à un texte pour obtenir une bonne contraction. Voici une description de ces approches, passées en revue dans [YMP06].

1.2.1 L'extraction de phrases

Ici encore, il s'agit d'extraire des éléments saillants du texte, mais ce sont des phrases complètes qui sont extraites, et non des syntagmes. Elles sont ensuite regroupées en un texte résumé de type *extract*. L'analyse sémantique n'a pas ici besoin d'être particulièrement précise et fine et le risque de perte d'information majeure³ est réduit. Cependant, il se peut qu'une phrase extrêmement longue contenant un élément saillant parmi de nombreux éléments "inutiles" soit conservée dans le texte contracté. Par exemple, l'extraction de phrase appliquée au court texte (1), dans un contexte où l'élément saillant serait l'autoroute conserverait la première phrase en négligeant la seconde.

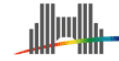
1. *La longue file ininterrompue de voitures se tortillait lentement le long de l'autoroute surchargée. Personne n'avancait.*

1.2.2 L'extraction de constituants

La phrase résumé

Cette méthode consiste à extraire d'un texte les éléments sémantiques saillants et à les regrouper et structurer en une phrase qui, on l'espère sera représentative du texte. Une analyse sémantique fine du texte est donc nécessaire, ainsi qu'un moyen de synthétiser la phrase résumé. Aucune perte majeure d'information n'est à exclure. Cependant, le taux de contraction est, bien entendu, très élevé.

³En un sens qui reste à définir.



La contraction de phrase

Cette méthode, sur laquelle j'ai travaillé pendant un mois et demi, permet de travailler, comme la précédente, au niveau du syntagme, mais en le plaçant dans la phrase au lieu de le considérer comme un élément du texte. Le principe de cette approche est de supprimer, dans chaque phrase du texte, les éléments qui ne sont pas syntaxiquement nécessaires (*i.e.* les compléments circonstanciels). Ainsi, le travail est principalement syntaxique et ne nécessite que des informations que la plupart des analyseurs syntaxiques fournissent. De plus, les considérations sémantiques qui y apparaissent sont minimales et servent uniquement à lever des ambiguïtés. Étudions plus en profondeur cette approche du résumé, et en particulier la manière dont elle a été développée à l'aide de SYGMART ([YMP05]).

1.3 Contraction de phrase

1.3.1 Présentation du système

Tout d'abord, la première approche choisie a été de déterminer les éléments de la phrase qui n'étaient pas nécessaires à la bonne formation syntaxique de la phrase. Ainsi, dans cette optique, tous les éléments que les grammaires traditionnelles considéraient comme effaçables (*c.f.* section 2.1) sont supprimés. Bien entendu, cette approche est acceptable lorsqu'il s'agit de contracter une phrase simple, mais peut donner lieu à des absurdités sémantiques dès lors que la phrase se complique ou est inscrite dans un contexte particulier.

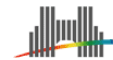
Le gros problème de la tâche de contraction consiste donc à conserver, en plus de la correction grammaticale, une certaine correction sémantique. Les différents obstacles à cette correction dans la première approche sont :

- les éléments repris par des anaphores, comme le montre l'exemple (2)⁴
- certains compléments circonstanciels, qui peuvent avoir une valeur de thème de la phrase, ou qui permettent parfois (notamment les circonstanciels de temps) de rendre cohérent un récit ou un texte écrit.

2. *Il leva les yeux vers sa mère. Celle-ci lui sourit.*

Une des principales difficultés rencontrées lors de la détection de tels phénomènes et de leur gestion est la complexité des langues naturelles ainsi que leur ambiguïté. Même un traitement qui tiendrait compte de toutes les informations sémantique directement disponibles à l'écrit ne pourrait lever certaines ambiguïtés, puisque l'humain même n'est parfois pas capable de le faire. Il s'agit donc ici de mettre au point des heuristiques s'approchant le plus possible d'un résultat "juste". Ainsi, en faisant une étude statistique des textes du corpus utilisé, on remarque que la règle générale pour les anaphores est qu'elles reprennent le syntagme accordé en genre et en nombre le plus proche à gauche dans le texte. Bien entendu, cette heuristique n'est utilisée qu'en cas d'ambiguïté et si aucune autre règle ne s'applique. Ceci n'est qu'un exemple des traitements appliqués au texte analysé, la réalité étant bien plus complexe, mais peu accessible au cours d'un stage court. Je continue donc par la présentation des résultats les plus marquants de ce système, présenté plus en détails dans

⁴Cet exemple (2) serait contracté, dans une première approche en : *Il leva les yeux. Celle-ci lui sourit.*



[YMP05, YMP06] et dont une version en ligne est disponible à l'adresse suivante : <http://www.lirmm.fr/~yousfi/compression.html>

1.3.2 Résultats

Les résultats sont ici fortement dépendants de la qualité de l'analyse syntaxique renvoyée par SYGFRAN. De plus, l'expérience montre que cette méthode est particulièrement efficace pour la contraction de textes narratifs⁵. Cependant, on remarque que l'exigence pure de complétude syntaxique ne suffit pas à garantir la conservation de la totalité des informations majeures⁶, comme le montre l'exemple (3), extrait de [Dod92].

3. *Enfin il se changea en pigeon vert. → Il se changea.*

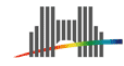
Comme on peut le voir en utilisant la version disponible en ligne du logiciel, le système conserve intacts les résultats de l'analyse syntaxique et ne les supprime que lors de l'application d'OPALE, dernière étape correspondant souvent à un affichage à l'écran. Ainsi, un autre ensemble de règle de marquage pourrait accorder plus d'importance aux adjectifs épithètes et négliger totalement les compléments d'agent ou les circonstants de temps et donner un autre résultat, suivant le type de texte d'entrée. Le système tient déjà compte de certaines informations sémantiques qui permettent notamment de traiter les cas cités ci-dessus. Il s'agira donc d'effectuer une amélioration du système existant pour tenir compte d'informations lexico-syntaxiques plus importantes, fournies par l'utilisateur ou obtenues de manière automatique par un autre moyen afin d'éviter des problèmes d'ambiguïté sémantique tels que celui rencontré dans l'exemple (3), ainsi que pour améliorer le traitement dans certains cas où des éléments non traités actuellement sont facultatifs (notamment les compléments d'objet). Nous voici donc en mesure de préciser le sujet du stage à la lumière du contexte du travail existant.

⁵L'importance des éléments facultatifs n'est, dans ce cas, que purement descriptive et ils n'apportent souvent que peu d'informations utiles à la compréhension. Dans un article scientifique, par contre, les adjectifs épithètes ne pourront que rarement être supprimés (*Une application surjective*).

⁶Ici, une "information majeure" est une information qui, si elle est supprimée change radicalement le sens de la phrase ou fait perdre sa cohérence au texte.

Problématique du stage

Le but de mon travail au cours de ce stage est donc de trouver une méthode permettant, à partir d'une ressource lexicale à définir (existante ou à constituer), de déterminer quels éléments sont essentiels à la complétude syntaxique, puis sémantique, de la phrase et de les marquer comme tels. Ce marquage pourra se faire en amont d'un traitement plus général concernant la contraction et ses résultats doivent être utilisables aisément dans un tel contexte. Un autre objectif est d'effectuer ce travail en utilisant le minimum d'informations sémantiques et en se contentant d'informations morpho-syntaxiques afin de pouvoir l'utiliser avec un système tel que SYGFRAN. Bien entendu, toute autre approche est possible. La recherche d'un lexique approprié, ou sa constitution, nécessitera d'étudier les théories grammaticales et linguistiques récentes, afin de déterminer les informations qu'il faudra y intégrer pour pouvoir effectuer le marquage des compléments essentiels. Enfin, il sera peut-être utile d'effectuer un relevé des différents cas d'ambiguïtés pouvant se présenter et qui ne seraient pas traités (cas marginaux, cas très difficiles, voire impossibles sans informations sémantiques...) et de proposer des solutions, si possible, aux problèmes qui pourraient se poser.



Chapitre 2

Considérations Grammaticales

Afin de pouvoir résoudre ce problème, il est nécessaire de s'intéresser de plus près à ce que linguistes et grammairiens pensent des compléments, qu'ils soient d'objets ou circonstanciels, puisque ceux-ci sont au centre de nos préoccupations. Nous verrons d'abord, pour nous rafraîchir la mémoire, ce que disent les grammaires dites traditionnelles, souvent plus intuitives, puis nous nous intéresserons à un type de grammaires apparues plus récemment : les grammaires d'unification, plus formelles, qui sont apparues avec les exigences liées au traitement automatique des langues naturelles.

2.1 Grammaires traditionnelles

Pour de nombreux grammairiens, les "compléments circonstanciels" traditionnels sont en réalité à répartir dans deux catégories de compléments (comprenant les compléments d'objet) : les compléments essentiels, liés au verbe (dont font partie les compléments d'objet), et les compléments facultatifs, aussi appelés "circonstants" et qui sont liés à la phrase ([Tom01]). La première difficulté à surmonter est donc de distinguer compléments essentiels et compléments facultatifs. Pour cela, nous disposons de l'approche traditionnelle, qui caractérise le complément essentiel par ses propriétés grammaticales. Comme nous l'apprenons dans [Tom01, GT04] : le complément du verbe n'est ni déplaçable (4), ni effaçable (5)¹, et est facilement remplaçable par un pronom (6).

4. *Paul mange une pomme.* → **Une pomme Paul mange.*

5. *Paul dort son dernier sommeil.* → *Paul dort.*

6. *Elle vit sa vie.* → *(Sa vie,) elle la vit.*

Les circonstants, quant à eux, sont déplaçables (7), effaçables (8), multipliables à volonté (9), et ne sont pas pronominalisables (10)², ce qui permet de les distinguer rapidement des compléments essentiels, comme il est écrit dans [RPR04].

7. *En août, les jours commencent à raccourcir.* → *Les jours commencent à raccourcir en août.*

8. *Cette année, l'été a été pluvieux.* → **L'été a été pluvieux.*

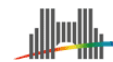
9. *Cette année, en Alsace, contrairement aux prévisions de la météo, l'été a été pluvieux, au grand dam des vignerons.*

10. *Pierre a éternué pendant le discours du président.*

11. *Pierre a éternué dans cet hôtel.* → *(Cet hôtel,) Pierre y a éternué.*

¹Ici, la grammaticalité est conservée, mais une nuance importante est perdue

²Sauf dans un nombre restreint de cas, notamment les circonstants de lieu (11)



De plus, cette approche propose des méthodes pour différencier aisément les divers type de compléments du verbe. En particulier, il peut être intéressant de savoir identifier les compléments d'objet (4) et les attributs du sujet (12), qui sont souvent obligatoires et ne permettent en général de ne supprimer que quelques mots.

12. *Jérémy est maladroit.*

L'inconvénient majeur d'une telle approche dans le contexte qui nous intéresse est la difficulté de l'automatiser. En effet, comment tester de manière algorithmique des notions de bonne formation sémantique ? Il faut donc, pour pouvoir analyser, et résumer, un texte de manière automatique, mettre au point une nouvelle approche de la syntaxe. C'est ce qui a été fait par les linguistes avec le développement des grammaires d'unification.

2.2 Grammaires d'unification

Les grammaires d'unification se sont développées en marge du modèle syntaxique traditionnel "chomskyen" pour d'une part, unifier l'étude de la syntaxe avec celle du lexique et de la sémantique, et d'autre part, permettre, par une formalisation plus explicite, des applications en traitement automatique des langues. [Abe93] présente assez bien cette famille de grammaires, en explicitant les plus répandues et en les appliquant en partie à la langue française. Voici une courte présentation de l'une d'entre elles, plus particulièrement adaptée au travail à fournir ici.

2.2.1 *Lexical functional grammar* (LFG)

La grammaire lexicale fonctionnelle (*lexical functional grammar* ou *LFG*), a été définie à la fin des années 70 par J. BRESNAN et R. KAPLAN dans [Bre82]. Ce modèle est moins complexe que les grammaires génératives transformationnelles utilisées jusqu'alors³. L'idée est de décrire la phrase non seulement au niveau de ses dépendances syntaxiques, représentées par la structure de constituants (ou structure *c*), mais aussi au niveau des dépendances lexicales, représentées par une structure de traits, dite structure fonctionnelle (ou structure *f*). Le résultat est une structure arborescente dont les noeuds sont étiquetés non seulement par les informations syntaxiques, mais aussi par des informations de dépendance lexicale.

Les figures 2.1 et 2.2, tirées de [Abe93], montrent l'utilité linguistique de cette représentation à deux niveaux. En effet, deux phrases sémantiquement proches ont une représentation arborescente (structure *c*, sur la figure 2.1) très différentes l'une de l'autre, mais leurs structures *f* (sur la figure 2.2) sont très proches. On remarque déjà dans ces structures *f* simplifiées un point central des grammaires d'unification, et en particulier de LFG : l'organisation à base de prédicats (Pred), instaurant une hiérarchie dans la phrase ou le syntagme, gouverné par une tête dont dépendent les autres syntagmes. Ainsi, à une phrase bien formée correspondront une structure *c*, obtenue par dérivation de règles de grammaire hors-contexte dotées d'une description fonctionnelle⁴, et une structure fonctionnelle, définie comme la structure de traits mini-

³LFG est, dans le pire cas, équivalent à une grammaire contextuelle, contrairement aux grammaires génératives à la Chomsky

⁴Un ensemble d'équations décrivant la structure fonctionnelle normalement associée à l'arbre de dérivation.

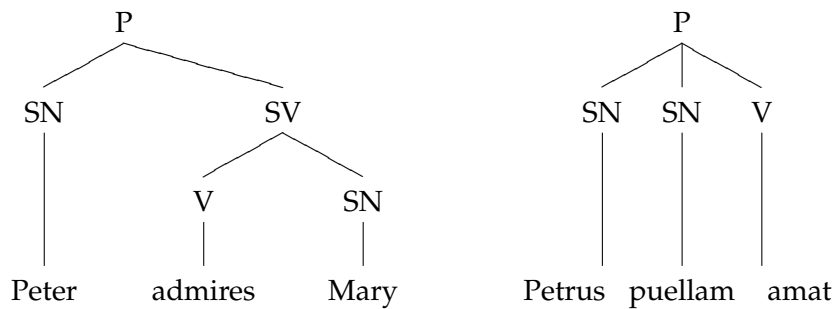
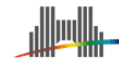


FIG. 2.1 – Structures c de phrases sémantiquement proches en anglais et en latin

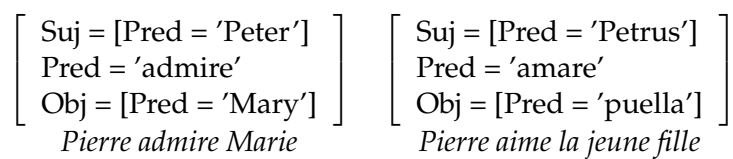


FIG. 2.2 – Structures f correspondantes

male satisfaisant la description fonctionnelle. C'est dans la recherche de cette structure fonctionnelle qu'intervient l'unification. La figure 2.3 représente l'unification d'un ensemble de règles. Un échec de l'unification signifierait que la phrase n'est pas bien formée au vu des règles fournies.

Les structures f un peu plus complètes représentées sur la figure 2.3 contiennent des informations dites de "sous-catégorisation" à la suite du prédicat, indiquant les syntagmes imposés par le prédicat (la tête du groupe). Ainsi, dans l'exemple, "dort" ne demande qu'un sujet.

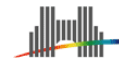
L'unification permet donc d'étiqueter un arbre syntaxique et d'en extraire la structure f correspondante à la phrase qu'il représente, à la condition de disposer des informations concernant la description fonctionnelle, et particulièrement la sous-catégorisation, de chacun de ses noeuds. En particulier, il est très important de connaître, en général et plus spécialement pour effectuer la reconnaissance d'éléments importants de la phrase, la sous-catégorisation des verbes, qui gouvernent souvent les syntagmes où ils se trouvent.

2.2.2 Le Lexique-Grammaire

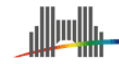
Établir la sous-catégorisation de tous les verbes d'une langue naturelle, dans tous leurs emplois et toutes leurs acceptions est un travail de titan. Cependant, Maurice GROSS et son équipe ont commencé, dans [Gro75], à écrire des tables représentant la sous-catégorisation des verbes courants du français, regroupés par type de construction.

De nombreux travaux en TAL, notamment les travaux de la communauté INTEX, s'appuient sur les résultats de GROSS, qui est sûrement la base de données lexicale la plus exhaustive à ce jour en ce qui concerne les traits de sous-catégorisation.

Cette ressource pourra donc être utilisée par la suite pour la constitution d'une res-



source lexicale, mais est aussi intéressante d'un point de vue plus théorique pour mieux comprendre le fonctionnement global des grammaires d'unification et avoir un aperçu plus complet du problème.



Chapitre 3

Marquage des compléments

3.1 Principes et choix théoriques

La réalisation d'un système répondant au problème posé passe, comme nous l'avons indiqué plus haut par l'utilisation ou la construction d'une ressource lexicale adaptée à l'approche choisie. Nous verrons d'abord ce qui a orienté ce choix, puis comment ce choix a lui-même influencé la création du lexique.

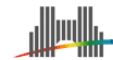
3.1.1 Choix d'une approche particulière

J'ai tout d'abord approché le travail à effectuer d'un point de vue très général, en envisageant de créer un système capable, à partir d'une ressource lexicale, de produire automatiquement un ensemble de règles TELESi correspondant aux informations contenues dans le lexique. J'ai beaucoup avancé dans cette direction, en tentant de structurer les différentes constructions lexico-syntaxiques et d'établir un ordre partiel sur l'ensemble les contenant. Il suffirait alors, lorsqu'un verbe est rencontré, de remplacer la construction C où il apparaît par une construction C' minimale vis-à-vis de l'ordre imposé telle que C' est plus petite que C dans ce même ordre. C'est en échangeant à propos des différentes définitions possibles d'une telle relation d'ordre avec Augusta MELA qu'il est apparu qu'une telle opération, menée de manière statique, s'approchait, sans atteindre leur efficacité, des différentes méthodes afférentes aux grammaires d'unification. Le principal inconvénient de cette méthode était son aspect de compilation statique d'une ressource lexicale, qui empêchait d'inclure au traitement des informations sémantiques qui pourraient être extraites de manière dynamique. De plus, il est très difficile d'écrire (en particulier de manière automatique) des règles TELESi cohérentes. Au fil des différentes réunions qui ont suivi, mon travail s'est donc orienté plus particulièrement vers l'étude des grammaires d'unification et vers leurs possibilités d'utilisation dans le contexte du stage. Il m'a donc fallu étudier un corpus (composé du conte polynésien présenté en annexe A.1 et des deux premiers chapitres de [dSE43]) afin de déterminer quelles informations sur les verbes étaient nécessaires pour distinguer les compléments essentiels des compléments facultatifs.

3.1.2 Constitution du lexique

Voici une description de ce que devra contenir la ressource lexicale qui sera finalement utilisée, à la lumière des considérations grammaticales du chapitre 2.

Pour décrire le contenu de cette ressource, la figure 3.1 représente quelques entrées du lexique "idéal", écrites dans un langage non-spécifique, qui se veut le plus explicite



possible. La ressource elle-même devra bien entendu tenir compte des formats d'entrée attendus par le système utilisé pour effectuer l'analyse, ainsi que de celui utilisé pour effectuer le marquage, qui pourraient éventuellement être tous deux différents de SYGMART.

Légende et explication de la figure 3.1

Le champ LEMME contient le lemme verbal, sa racine infinitive. Il est utile, de manière évidente, de connaître le verbe auxquelles s'appliquent les informations données dans la suite du tableau.

Le champ FORME donne des informations précises sur la forme verbale, qui peut fortement influencer l'importance des différents éléments de la phrase¹. En particulier, il indique les formes passives, impersonnelles, et réfléchies. Il sera donc aussi nécessaire de détecter, dans certains cas où SYGFRAN n'est pas efficace, ces différentes voix/constructions.

13. *Jean a lavé le sol.* → <Suj :SN ;Obj :SN,SCOMP>

14. *Jean s'est lavé (∅/le corps/les mains).* → <Suj :SN ;(Obj :SN[partie du corps])>

Il peut prendre trois valeurs ou toute combinaison de ces trois valeurs : *Pas* pour les voix passives, *Imp* pour les constructions impersonnelles, et *Ref* pour les constructions réflexives.

Le champ AUXILIAIRE contient l'auxiliaire avec lequel se construisent les temps composés du verbe. En effet, celui peut varier suivant la voix, l'acception ou d'autres facteurs (comme illustré dans les phrases (15) et (16)). *A priori*, cette information n'est que peu utile, mais elle peut peut-être mener à une diminution du nombre de cas d'ambiguïtés suivant le corpus. Ce champ n'a que deux valeurs possibles : *être* ou *avoir*. Il est envisageable, mais je ne l'ai pas reconstruit dans le corpus, qu'un verbe puisse être construit avec les deux auxiliaires indifféremment.

15. *Il s'est changé trois fois aujourd'hui.*

16. *Il a beaucoup changé ces derniers temps.*

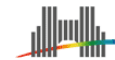
Pour comprendre les deux champs suivants, il nous faut savoir ce qu'est un verbe support. Dans [RPR04, VII-1.4.8, p. 232], les verbes supports «sont des verbes qui, à côté de leurs emplois ordinaires, peuvent se combiner avec un nom, un adjectif ou un groupe prépositionnel pour construire une forme complexe fonctionnellement équivalente à un verbe». De plus, cette forme complexe a ses propres caractéristiques, ainsi qu'un sens différent, ce qui justifie une entrée séparée dans le lexique.

L'étude du corpus nous montre qu'il est nécessaire, pour analyser et traiter correctement les verbes supports, de connaître à la fois la catégorie grammaticale de l'élément supporté (CATSUP), et, bien entendu, son lemme (LEMSUP). En effet, le lemme est nécessaire, mais quelques exemples ((17) et (18), par exemple), rares, mais importants car ils ont d'autres conséquences, rendent nécessaire la connaissance de la catégorie.

17. "Faire le beau", CATSUP=N, LEMSUP="beau", SUBCAT=<Suj :SN>

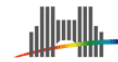
18. "Faire beau", FORME=Imp, CATSUP=Adj, LEMSUP="beau", SUBCAT=<Suj :II>

¹Les phrases (13) et (14) montrent l'importance, par exemple, de détecter une forme réfléchie qui modifie la sous-catégorisation du verbe *laver*.



LEMME	FORME	AUXILIAIRE	CATSUP	LEMSUP	SOUS-CATÉGORISATION (SUBCAT)	PROPRIÉTÉS
partir	∅	être	Prep(à)	recherche	<Suj :SN,SINF,SCOMP ;Obj :de-SN>	∅
être	∅	avoir	∅	∅	<Suj :SN,SINF,SCOMP ;AttSuj :SN,Adj>	∅
manger	∅	avoir	∅	∅	<Suj :SN ;Obj :SN,SCOMP>	∅
faire	∅	avoir	∅	∅	<Suj :SN,SINF,SCOMP ;Obj :SN>	∅
faire	Impersonnel	avoir	∅	∅	<Suj : "I" ;Obj :SN-météo,Adj-météo>	∅
faire	∅	avoir	N	attention	<Suj :SN,SINF,SCOMP ;(Obj :à-SN,à-SINF)>	∅
faire	∅	avoir	N	expérience	<Suj :SN,SINF,SCOMP ;(Obj1 :de-SN ;Obj2 :sur-SN)>	∅
faire	∅	avoir	N	connaissance	<Suj :SN ;Obj :de-SN>	∅
gronder	∅	avoir	∅	∅	<Suj :SN ;Obj :SN-humain>	∅
protester	∅	avoir	∅	∅	<Suj :SN,SCOMP>	∅
moquer	Réflexif	être	∅	∅	<Suj :SN,SCOMP ;Obj :de-SN>	∅
mettre	∅	être	∅	∅	<Suj :SN,SCOMP ;Obj :à-SINF>	CTRLSUIJ
pousser	∅	avoir	∅	∅	<Suj :SN ;Obj1 :SN ;Obj2 :à-SINF>	CTRLOBJ

FIG. 3.1 – Extrait du lexique



Le premier a trois valeurs possibles, correspondant aux trois catégories possibles, d'après Riegel *et al.* : *N* pour un nom ou un syntagme nominal, *Adj* pour un adjectif, et *Prep* pour un groupe prépositionnel.

Le champ SUBCAT contient les informations de sous-catégorisation proprement dites, à savoir les compléments imposés par le verbe. Plusieurs informations sont apportées pour chaque complément. Tout d'abord sa fonction syntaxique (Sujet (Suj), Complément d'Objet (Obj), Attribut (Att)...), puis les catégories grammaticales possibles pour chaque fonction (syntagme nominal (N), adjectif (Adj), adverbe (Adv), subordonnée complétive (SCOMP)...) et enfin, des informations plus précises sur la catégorie sémantique du complément (humain, partie du corps, animé, comestible...). Les différentes valeurs possibles sont susceptibles d'être modifiées suivant le corpus étudié, et il serait long et inutile d'en faire ici un inventaire complet. Cependant, une étude complète serait bien entendu nécessaire pour être en mesure de fournir un système opérationnel, ne serait-ce que sur un petit ensemble de textes.

Cette sous-catégorisation présente la particularité de contenir aussi des éléments essentiels "facultatifs". Ces éléments apportent des précisions qui, si elles sont supprimées, peuvent engendrer une ambiguïté sémantique. Ainsi, ces éléments, présentés entre parenthèses dans le lexique, ne sont pas obligatoires, mais ne doivent pas être supprimés, au moins dans une première approche, s'ils sont présents.

Le dernier champ, PROPRIÉTÉS, contient des informations supplémentaires, relatives à la construction de la phrase elle-même, et permettant d'affiner l'analyse en constituants. En particulier, on y mettra des informations relatives à des formes passives irrégulières, ou plus souvent les informations concernant l'éventuel accord (contrôle) d'une proposition infinitive complément². Une proposition infinitive peut aussi être impersonnelle et transmettre cette caractéristique à la principale, avec certains verbes, comme dans l'exemple (21). Cette information sera aussi donnée par ce champ.

19. *Je dois aller me laver.*

20. *Je lui ai dit d'aller se laver.*

21. *Il devrait pleuvoir.*

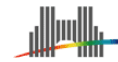
Au cours de mes recherches, il s'est avéré qu'aucune ressource lexicale existante ne fournissait la totalité de ces informations. Il sera donc nécessaire de construire cette ressource au fur et à mesure des expérimentations, en tenant compte du fait qu'une entrée représente une acception d'un verbe et non son lemme, puisque certains verbes peuvent avoir plusieurs acceptions possédant des caractéristiques distinctes³.

3.2 Réalisation du marquage

Disposer de ces informations ne suffit pas, en soi-même à effectuer l'opération souhaitée. Encore faut-il pouvoir, et savoir les utiliser. Je décrirai ici les règles les plus générales possibles, tout en sachant, et en précisant, qu'une réalisation pratique sous forme d'une grammaire TELESY devra bien entendu tenir compte de l'analyse SYG-FRAN sur l'entrée et gérer de nombreux cas particuliers (pronominalisations, flexions

²Le contrôle de l'infinitif est l'élément qui sert de sujet à la proposition. Ce contrôle peut-être le sujet de la phrase principale (ex. (19)), ou son complément d'objet ((20)).

³la principale conséquence est une ambiguïté qui apparaît à la lecture du dictionnaire, puisque celle-ci peut renvoyer plusieurs entrées si la sélection ne s'effectue que sur le lemme



des déterminants...). Plusieurs approches étaient possibles, ici encore. J'ai choisi ici un approche en couches successives permettant encore un raffinement des critères de sélection et de traitement à différents niveaux.

La première couche (3.2.1) se chargerait de "corriger" l'analyse SYGFRAN en remarquant bien les constructions particulières (formes impersonnelles, formes réfléchies et voix passives), ce qui, en plus de rendre l'analyse plus pertinente, peut permettre de réduire fortement le nombre d'entrées à considérer dans le dictionnaire en affinant le filtrage.

Une deuxième couche (3.2.2) se chargerait de répertorier, pour toutes les formes verbales, l'ensemble des syntagmes potentiellement sous-catégorisés (tous les groupes syntaxiques qui pourraient être compléments de ce verbe).

La troisième couche (3.2.3) effectuerait alors le traitement proprement dit, considérant chaque complément potentiel et vérifiant sa présence dans les entrées du lexique correspondantes. Si le groupe est présent dans toutes les entrées avec la même fonction, on lui affecte cette fonction. Sinon, s'il existe une entrée où ce groupe est présent, on divise l'arbre comme sur l'exemple d'analyse ambiguë 1.3 et on continue l'analyse. Dans le cas où aucune des entrées lexicales ne sous-catégorise un élément de la nature souhaitée, alors ce groupe est un complément de phrase ou dépend d'un autre verbe dans la phrase.

Enfin, une quatrième couche permettrait de remettre en forme l'arbre d'analyse syntaxique en plaçant les compléments du verbe sous le groupe verbal.

3.2.1 Pré-traitement : Mise en forme de l'analyse SYGFRAN

Formes réfléchies

Pour ce point, au moins dans les cas généraux, SYGFRAN détecte bien les emplois réfléchis des verbes. Le groupe verbal possède alors l'étiquette VOIX= (PRONOM). Dans le cas où l'expérience mettrait en évidence une défaillance de SYGFRAN pour mettre en évidence les emplois réflexifs, il est possible d'écrire un ensemble de règles simples analysant la personne et le nombre du sujet et vérifiant qu'il s'accorde avec le pronom après avoir vérifié que celui-ci est sous une forme réfléchie (*me, se, nous, vous*). Si l'expérience montre à nouveau que cela ne suffit pas encore, on pourra marquer comme "potentiellement réfléchi" tout emploi d'un verbe avec un pronom sous forme réfléchie et lui appliquer à la fois les règles des formes réflexives et celles des formes classiques lors du traitement.

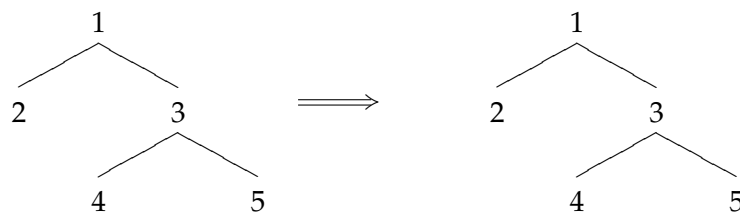
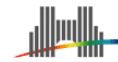
Formes impersonnelles

En dehors des cas où le verbe est purement impersonnel (verbes météorologiques, de survenance ou modalisateurs, par exemple (22), (23) et (24)), la détection des formes impersonnelles n'est pas aisée si on ne considère que des informations lexico-syntaxiques. En effet, la présence du pronom *il* en tant que sujet ne suffit pas, heureusement pour la communication et malheureusement pour l'analyse, à donner un sens impersonnel au verbe ((25) et (26)).

22. *Il pleut.*

23. *Quoi qu'il advienne.*

24. *Il faut manger tes épinards.*



K (3) =GV, CAT (4) =V,	TYP (4) ←IL
PROP (4) ⊃TRSF_IMP,	TYP (3) ←IL
K (5) =PHINF, TYP (5) ⊃IL,	
LEMME (2) = "il"	

FIG. 3.2 – Règle : propagation des formes impersonnelles infinitives

25. *Il/Jean est arrivé à 5 heures.*

26. *Il/*Jean est arrivé quelque chose.*

Tout d'abord, le caractère intrinsèquement impersonnel d'un verbe peut, dans certains cas et lorsqu'il est infinitif complément, être transmis au verbe principal (27). La règle la plus générale correspondante pourrait s'écrire 3.2. Ensuite, un verbe peut posséder plusieurs emplois dont un ou plusieurs sont impersonnels ((25) et (26)). Comme, dans ces cas, l'analyse dépend fortement du contexte, que nous tenons à écarter du prétraitement⁴, tout verbe dont le sujet est le pronom *il* sera marqué comme "potentiellement impersonnel" et recevra un double-traitement lors de la désambiguïsation. Parfois, le choix sera rapidement effectué au cours de l'étape de traitement proprement dite (3.2.3), par exemple, après la détection des formes supports (section 3.2.3) comme dans la phrase (28) où la détection de la forme support *faire*(adj. "beau") imposera la forme impersonnelle au verbe.

27. *Il devrait pleuvoir, aujourd'hui.*

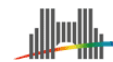
28. *Il fait beau.*

Voix passive

Une des principales difficultés de la détection des verbes à la voix passive est leur proximité structurale avec les temps composés. Il faut donc trouver des différences entre les constructions passives et actives composées d'un même verbe pour les distinguer. En particulier, la détection d'un complément d'agent introduit par la préposition *par* permet une détermination quasi-immédiate. En revanche, la distinction peut-être plus difficile à faire si les différences sont plus fines, voire purement sémantiques⁵. Le travail de détection immédiate des formes passives étant assez bien effectué par SYG-FRAN, il suffit de dédoubler l'arbre syntaxique dans certains cas ambigus (notamment

⁴Le prétraitement vise justement ici à extraire les informations immédiates

⁵Il s'agit ici de détecter les verbes à la voix passives et non les constructions à sens passif. Une telle détection pourrait être utile dans le cadre d'une reformulation ou d'une analyse sémantique, mais est ici inutile



(29), analysé comme l'auxiliaire *être* suivi de son attribut du sujet) pour avoir une analyse plus complète.

L'exemple (29) illustre très bien le propos car, suivant la manière dont on analyse la forme verbale *sont nourries*, le complément *au grain* est essentiel⁶ (auxiliaire + attribut du sujet) ou facultatif (passif, *au grain* peut même être considéré comme l'agent tout en restant facultatif).

29. *Ces poules sont nourries au grain.*

On pourrait inclure dans ce pré-traitement la détection des verbes supports, mais encore une fois, le but ici est de déduire autant d'informations que possible de l'analyse elle-même sans se servir de la ressource lexicale. Cette détection pourra donc se faire au cours de la troisième étape, où la ressource lexicale sera utilisée.

3.2.2 Repérage des compléments potentiels

Dans cette étape, il s'agit de faire l'inventaire de tous les groupes syntaxiques potentiellement rattachés à un même verbe. Ces groupes seront ensuite analysés à la lumière du lexique et retenus, ou non, en tant que compléments du verbe en question. Nous ne prendrons ici que les noeuds frères du groupe verbal étudié (pour retenir les sujets et les compléments essentiels qui seraient analysés par SYGFRAN comme des compléments circonstanciels), et ses descendants (pour retenir tous les groupes initialement analysés comme des compléments d'objet, compléments d'agent, attributs du sujet, attributs de l'objet...). On peut ensuite placer chacun des éléments retenus, sans plus de sélection, dans une étiquette du nœud correspondant au verbe. On pourra utiliser pour les identifier les numéros de nœud renvoyés par SYGFRAN lors de l'analyse. Une difficulté supplémentaire, due à l'imperfection de l'analyse SYGFRAN peut apparaître si le verbe considéré fait partie d'une subordonnée complétive dont les compléments sont mal analysés. Je n'ai à ce jour pas de solution à proposer, ces cas étant relativement rares dans les textes considérés. Une solution, trop coûteuse pour être envisageable serait de considérer tous les syntagmes de la phrase, mais je préfère me reposer sur l'hypothèse d'une analyse correcte et complète, qu'il est toujours possible d'envisager, quitte à rajouter un pré-traitement supplémentaire.

À la fin de cette étape, chaque verbe de la phrase dispose donc de la liste des syntagmes susceptibles d'être gouvernés par lui. Il "suffit" maintenant de parcourir cette liste en éliminant les syntagmes qui ne sont clairement pas compléments du verbe et en marquant ceux qui, après comparaison avec les entrées du lexique, pourraient en être.

3.2.3 Traitement : lecture du lexique et marquage des compléments

La première chose à faire au cours du traitement proprement dit est de repérer les constructions grammaticales les plus courantes qui n'ont pas été reconnues à la première étape (c.f. 3.2.1). Parmi ces constructions, les constructions à verbes supports ont une importance particulière puisque, comme nous l'avons vu, leur présence entraîne d'importantes modifications de la sous-catégorisation du verbe et de ses caractéristiques.

⁶d'un point de vue sémantique, la syntaxe se fichant pas mal du droit des animaux

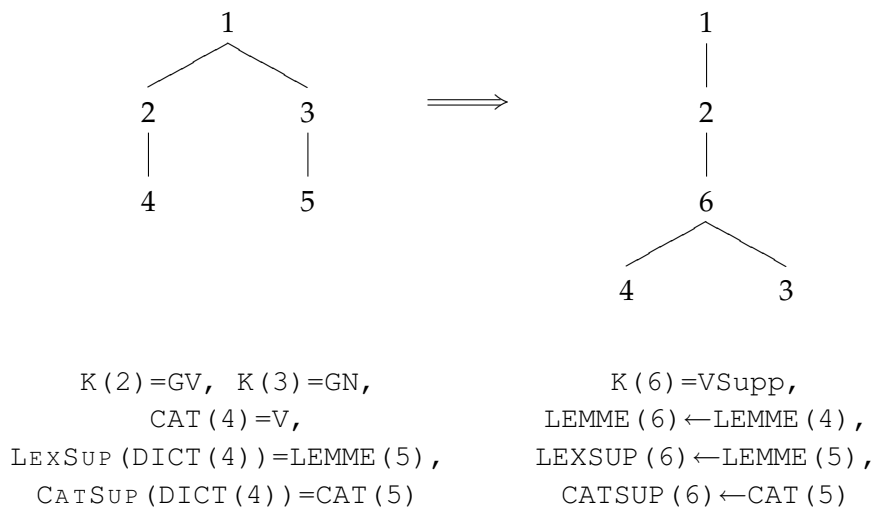
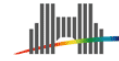


FIG. 3.3 – Règle : traitement des verbes supports

Détection des formes supports

Les constructions à verbes supports sont elles aussi assez difficiles à détecter bien qu'elles ne diffèrent que peu des formes figées qui, elles, sont faciles à repérer. En effet, les formes supports peuvent subir des flexions (30), voire des inversions ou des anaphores (31), comme toute autre construction non-figée. Cependant, il est nécessaire, même dans ces cas, de les détecter, puisque la sous-catégorisation du verbe est modifiée.

30. *Jean a de l'admiration pour Marie./Jean a une certaine admiration pour Marie.*

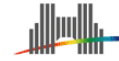
31. *Le jugement que Jean porte sur ton livre est positif.*

Je ne traiterai pas ici le cas des anaphores, mais des heuristiques permettraient de savoir avec une bonne probabilité à quel élément se rapporte la conjonction ou le pronom et de se ramener au cas étudié ci-dessous.

Pour détecter une forme support, il faut d'abord trouver la lexie éventuellement supportée par le verbe. Une fois celle-ci détectée, on place dans le groupe verbal, sous un nouveau noeud VSUPP contenant le lemme verbal, l'ensemble du syntagme la contenant. Cela donnerait une règle du type de la figure 3.3, où $DICT(x)$ est une lecture du dictionnaire avec le filtre formé par les étiquettes du noeud x .

Marquage effectif

Au cours de ce traitement, il faut, pour chaque verbe, parcourir le trait de sous-catégorisation des différentes entrées du lexique correspondantes à la forme détectée (le filtrage peut se faire sur le LEMME, la FORME et l'AUXILIAIRE, puisque la première étape (3.2.1) a permis de les déterminer avec précision. À chaque fonction sous-catégorisée par le verbe, on essaye ensuite d'associer un des syntagmes détectés lors de la deuxième étape du prétraitement (3.2.2). Si plusieurs éléments peuvent remplir une même fonction, on peut soit dédoubler l'arbre d'analyse, soit mettre en place des algorithmes approchés permettant d'en sélectionner un unique. Enfin, les seules acceptions du verbe possibles sont celles qui ont reconnu tous leurs compléments. S'il



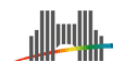
en reste plusieurs, on peut procéder de même, soit à une division de l'arbre, soit à une sélection grâce à des heuristiques approchées. La première solution reste à privilégier dans ce cas, puisque rien n'assure qu'un des syntagmes choisis n'est pas en réalité complément d'un autre verbe. Dédoubler l'arbre d'analyse permettrait alors d'effectuer un traitement plus exhaustif, autorisant les impasses, qui seront élaguées par la suite.

C'est au cours de cette étape, plus particulièrement, que pourront être utilisées les méthodes issues des grammaires d'unification. En effet, si on associe à la sous-catégorisation des équations fonctionnelles, comme dans LFG (*c.f.* 2.2), la sélection des compléments se fait de manière beaucoup plus fine (en particulier en rendant leur importance aux informations morphologiques de genre et de nombre), laissant la place à moins d'ambiguïtés.

3.3 Résultats et développements futurs

La durée du stage et la complexité du problème ne m'ont pas permis d'arriver à un stade de développement suffisamment avancé pour produire un résultat utilisable. De plus, l'intégration d'un système général dans la chaîne SYGFRAN existante, construite à bases d'expériences particulières ne se fera que difficilement, si elle se fait. Cependant on peut déjà dire que, si un tel système permettra une plus grande cohérence de la compression des phrases, il en réduira sensiblement l'efficacité en termes de taux de compression. Enfin, les résultats seront toujours imparfaits puisque certains phénomènes, même courants, de la langue ne sont pas traités.

Pourtant, même si je n'ai pas atteint un stade suffisant pour produire un système fonctionnel, je pense avoir atteint une bonne compréhension générale des problèmes inhérents au TAL et à la contraction de phrase. Ainsi, mon étude a permis à l'équipe de faire, ou refaire, certaines constatations sur les choix initiaux de la conception de SYGFRAN et m'a permis de prendre conscience de l'ampleur du travail à fournir en TAL, autant en termes de travail théorique général (notamment dans le domaine des grammaires formelles) qu'en termes d'étude fine de cas particuliers récurrents (problèmes de linguistique).

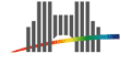


Conclusion

Les travaux que j'ai effectués durant ce court stage, même s'ils n'ont pas mené à des résultats directement utilisables, auront permis, je pense, aux membres de l'équipe TAL de revoir SYGMART, non comme un système expérimental sur lequel fournir un travail pratique permettant d'en prouver l'efficacité, mais aussi comme un système suffisamment proche des nouvelles théories de la grammaire pour qu'elles puissent y être adaptées efficacement.

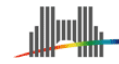
Le travail fourni au cours de ce mois et demi mérite d'être approfondi, non seulement pour son intérêt pratique et ses applications, mais aussi pour les questions théoriques qu'il suscite, en particulier sur la compatibilité du système de transformations que représente SYGMART avec les grammaires d'unification (LFG ou une autre). Pour rendre utilisables et testables les idées qui ont émergé de mon travail, une formalisation des règles et leur écriture au format TELES, ainsi que la constitution d'un lexique plus complet grâce à l'étude d'un corpus plus étendu seront nécessaires.

J'espère pouvoir, pour ma part, travailler à nouveau, si ce n'est sur SYGMART même, dans un domaine proche du TAL, car ces premiers pas dans le monde de la recherche, et dans un domaine aussi complexe que le traitement des langues naturelles, m'ont persuadé du travail phénoménal qu'il reste à accomplir dans ces domaines nouveaux, aux frontières de plusieurs disciplines.



Liste des annexes

A	Extrait d'un conte polynésien	33
A.1	Texte Initial	33
A.2	Résultat de la compression	35



Annexe A

Extrait d'un conte polynésien : *La Légende de Maui* [Dod92]

A.1 Texte Initial

Maui part à la recherche de ses parents.

A partir de ce soir-là, Maui fut le favori de sa mère : même s'il faisait des bêtises, elle ne le grondait pas. Quand ses frères protestaient, il se moquait d'eux parce qu'il savait avoir la protection de sa mère. Mais pendant son absence, il devait faire attention à ne pas dépasser les limites, sinon il risquait d'être puni par eux au cours de la journée.

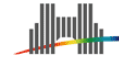
Une nuit, Maui imagina un tour à jouer à sa mère afin de découvrir où elle allait. Une fois tous les autres endormis sur leurs nattes, il se releva et fit le tour de la maison, examinant les grands stores tressés qui la fermaient pour la nuit. Partout où filtrait la clarté d'une étoile, il bouchait vite l'ouverture avec des étoffes d'écorce et calfeutrait même les fentes avec des roseaux. Puis il déroba le manteau, la ceinture et la couronne de sa mère et les cacha en se disant qu'il en aurait besoin plus tard.

Maui reprit alors sa place sur les nattes et décida de rester éveillé. La longue nuit passa lentement sans que sa mère ne bouge. Quand vint le matin, pas un rai de lumière ne put percer pour éveiller les dormeurs. Bientôt ce fut l'heure où le soleil grimpait au-dessus de l'horizon. D'habitude Maui pouvait distinguer dans la pénombre les formes des pieds de ses frères à l'autre bout de la maison, mais ce matin il faisait trop noir. Et sa mère continuait à dormir.

Au bout d'un moment elle bougea et marmonna : « Quelle sorte de nuit est-ce donc pour durer si longtemps ? » Mais elle se rendormit parce qu'il faisait aussi noir qu'au cœur de la nuit dans la maison.

Finalement elle se réveilla en sursaut et se mit à chercher ses vêtements. Courant de tous côtés, elle arracha ce que Maui avait fourré dans les fentes. Mais c'était le jour ! Le grand jour ! Le soleil était déjà haut dans le ciel ! Elle s'empara d'un morceau de tapa pour se couvrir et se sauva de la maison, en pleurant à la pensée d'avoir été ainsi trompée par ses propres enfants.

Sa mère partie, Maui bondit près du store qui se balançait encore de son passage et regarda par l'ouverture. Il vit qu'elle était déjà loin, sur la première pente de la montagne. Puis elle s'arrêta, saisit à pleines mains



un arbuste de *tiare* Tahiti, le souleva d'un coup : un trou apparut, elle s'y engouffra et remit le buisson en place comme avant.

Maui jaillit de la maison aussi vite qu'il put, escalada la pente abrupte, trébuchant et tombant sur les mains car il gardait les yeux fixés sur l'arbuste de *tiare*. Il l'atteignit finalement, le souleva et découvrit une belle caverne spacieuse qui s'enfonçait dans la montagne.

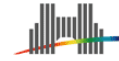
Plus tard, tandis que ses frères étaient très affairés à se baigner dans le frais ruisseau et à chercher un fruit de l'arbre à pain à se mettre sous la dent, Maui les questionna : «Où croyez-vous donc que notre père et notre mère passent la journée ?»

«Comment le saurions-nous ? Et pourquoi te tracasses-tu avec ça ? Tu ne peux donc pas vivre tranquillement avec nous ? Que nous importe notre père ou notre mère ? Est-ce qu'elle nous a élevés avec de la bonne nourriture ? Pas du tout : elle était toujours partie. Peut-être bien que le grand *Ta'aroa*, dieu du ciel, est notre père et qu'il a envoyé ses enfants ici-bas pour s'occuper de nous ! *Niu-Hiti*, la douce brise qui rafraîchit la terre et les jeunes plantes ; *Hau-Ri'i*, le vent humide qui les mouille ; *Hau-Roto-Roto*, le beau temps qui les fait pousser ; *Tou-Ari'i*, le dieu de la pluie qui les arrose, et son frère, *Ti-Ari'i*, qui les nourrit de ses rosées. *Ta'aroa* a envoyé toute sa famille pour permettre à notre nourriture de pousser. Ensuite *Papa*, la grande déesse mère de la terre, a fait germer ses graines pour nous tous qui sommes ses enfants.»

«Mais oui, bien sûr», leur répondit le petit Maui, «tout ce que vous dites est vrai. C'est même encore plus vrai pour moi que pour vous, parce que c'est la mer qui a été ma nourrice et ses bouillonnements d'écume mon lait. Vous, vous avez été nourris au lait de notre mère avant de pouvoir manger d'autres nourritures. Mais moi, ô mes frères, je n'ai jamais tété son sein, ni rien mangé de sa main. Et pourtant je l'aime pour l'unique raison qu'elle m'a porté en elle, et c'est parce que je l'aime que je souffre de ne pas savoir où elle se trouve avec mon père.»

Ses frères se sentirent surpris et charmés par le petit Maui quand ils l'entendirent parler de cette façon. Après avoir réfléchi un moment à ce qu'il avait dit, ils l'approuvèrent et l'encouragèrent à tenter de trouver leur père et leur mère.

Maui ne se tint plus de joie. Il se mit tout de suite à faire la magie dont il savait avoir besoin pour pénétrer dans la caverne sous l'arbuste de *tiare* et trouver son chemin souterrain dans l'autre monde. Il allait devoir voyager vite et il décida de se changer en oiseau. Il ne savait pas quel oiseau choisir. Il pensa bien sûr au *noha*, le pétrel qui niche dans un terrier de la montagne, mais il le jugea trop gros. Il se fit *maho*, marouette fuligineuse, mais ses frères pensèrent qu'il était trop petit et pas joli. Puis il devint *otaha*, grande frégate noire, mais ils trouvèrent effrayante cette créature aux ailes plus longues que leurs bras. Alors il essaya un oiseau après l'autre, le *'uriri*, petit chevalier voyageur à la voix claire et hardie, le *tarapapa*, hirondelle de mer bruyante à la voix gutturale, le *kivi*, courlis chasseur des petits crabes rouges, au long bec courbé comme un manche d'outil, le *otu'u*, aigrette sacrée qui tanguait sur ses hautes échasses, le *torea*, pluvier doré, puis le *a'o*, fou brun, et le *ua'ao*, fou à pieds rouges, trop comiques, et puis le *itata'e*



tout blanc et le *oa* tout brun, jusqu'à ce qu'il ait pris l'apparence de tous les oiseaux du monde, tour à tour. Enfin il se changea en pigeon vert.

Extrait de *La Légende de Maui*. [Dod92]

A.2 Résultat de la compression

Maui part à la recherche de ses parents.

A partir de ce soir-là, Maui fut le favori de sa mère : elle ne le grondait pas. Quand ses frères protestaient, il se moquait d'eux. Mais il devait faire attention à ne pas dépasser les limites, sinon il risquait d'être puni au cours de la journée.

Une nuit, Maui imagina un tour à jouer à sa mère. Une fois tous les autres endormis, il se releva et fit le tour de la maison. Partout où filtrait la clarté d'une étoile, il bouchait vite l'ouverture et calfeutrait même les fentes des roseaux. Puis il déroba le manteau, la ceinture et la couronne de sa mère et les cacha.

Maui reprit alors sa place et décida de rester éveillé. La longue nuit passa lentement. Quand vint le matin, pas un rai de lumière ne put percer. Bientôt ce fut l'heure. D'habitude Maui pouvait distinguer les formes des pieds de ses frères à l'autre bout de la maison, mais ce matin il faisait. Et sa mère continuait à dormir.

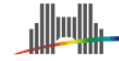
Au bout d'un moment elle bougea et marmonna : " Quelle sorte de nuit est donc-ce pour durer si longtemps ? Mais elle se rendormit.

Finalement elle se réveilla en sursaut et se mit à chercher ses vêtements. elle arracha ce Maui que avait fourré. Mais c'était le jour ! Le grand jour ! Le soleil était haut déjà ! Elle s'empara d'un morceau de tapa et se sauva.

Sa mère partie, Maui bondit près du store et regarda par l'ouverture. Il vit qu'elle était déjà loin. Puis elle s'arrêta, saisit un arbuste de tiare Tahiti, le souleva : un trou apparut, elle s'y engouffra et remit en place le buisson comme avant.

Maui jaillit aussi vite qu'il put, escalada la pente abrupte, trébuchant et tombant car il gardait les yeux fixés sur l'arbuste de tiare. Il l'atteignit finalement, le souleva et découvrit une caverne.

Plus tard, tandis que ses frères étaient très affairés à se baigner dans le frais ruisseau et à chercher un fruit de l'arbre à pain à se mettre sous la dent, Maui les questionna : " croyez Où que notre père et notre mère passent la journée-vous donc ? " Comment le saurions-nous ? Et pourquoi te tracasses-tu avec ça ? Tu ne peux donc pas vivre tranquillement avec nous ? Que nous importe notre père ou notre mère ? Est qu'elle nous a élevés avec de la bonne nourriture-ce ? Pas du du tout : elle était toujours partie. Peut-être bien que le grand Ta'aroa est notre père et qu'il a ses enfants envoyé ici-bas ! Niu-Hiti ; Hau-Ri'i ; Hau-Roto-Roto ; Tou-Ari'i et son frère. Ta'aroa a toute sa famille envoyé. Ensuite Papa a germer ses graines fait.-" Mais oui, bien sûr ", leur répondit le petit Maui, " tout ce vous que dites est vrai. C'est même encore plus pour moi que pour vous vrai. Vous,



vous avez été nourris. Mais moi, ô mes frères, je n'ai jamais son sein tété, ni rien mangé. Et pourtant je l'aime, et c'est parce que je l'aime que je souffre de ne pas savoir où elle se trouve ."

Ses frères se sentirent surpris et charmés. ils l'approuvèrent et l'encouragèrent à tenter de trouver leur père et leur mère.

Maui ne se tint plus de joie. Il se mit tout de suite à faire la magie. Il allait devoir voyager vite et il décida de se changer.

Il ne savait pas quel oiseau choisir. Il pensa bien sûr au noha mais il le jugea trop gros. Il se fit maho mais ses frères pensèrent qu'il était trop petit et pas joli. Puis il devint otaha mais ils trouvèrent cette créature aux ailes plus longues que leurs bras effrayante. Alors il essaya un oiseau après l'autre. Enfin il se changea.

Bibliographie

- [Abe93] Anne ABEILLÉ : *Les nouvelles syntaxes : grammaires d'unification et analyse du français*. Armand Colin, première édition, 1993.
- [Bre82] Joan BRESNAN : *The Mental representation of grammatical relations*. MIT Press, 1982.
- [Cha84] Jacques CHAUCHÉ : Un outil multidimensionnel de l'analyse du discours. *In Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pages 11–15, Morristown, NJ, USA, 1984. Association for Computational Linguistics.
- [Cha01] Jacques CHAUCHÉ : *SYGMART : manuel de référence*. LICIA, 11 bis, rue des Capucins. 92190 Meudon, 4.0 édition, avril 2001. <http://www.lirmm.fr/~chauche/REFERENCESYG/docsyg.html>.
- [Dod92] Edward DODD : *La Légende de Maui / Maui peu tini*. Haere Po, 1992.
- [dSE43] Antoine de SAINT EXUPÉRY : *Le Petit Prince*. 1943.
- [Gro75] Maurice GROSS : *Méthode en syntaxe. Régime des constructions complétives*. Hermann, première édition, 1975.
- [GT04] Joëlle GARDES-TAMINE : *La Grammaire*, volume 2. Syntaxe. Armand Colin, troisième édition, 2004.
- [RPR04] Martin RIEGEL, Jean-Christophe PELLAT et René RIOUL : *Grammaire méthodique du français*. PUF, troisième édition, février 2004. Deuxième tirage (juin 2005).
- [Tom01] Roberte TOMASSONE : À propos des « compléments circonstanciels ». *Les revues pédagogiques de la Mission laïque française Connaissance du français*, 43: 59–64, novembre 2001.
- [YMP05] Mehdi YOUSFI-MONOD et Violaine PRINCE : Utilisation de la structure morpho-syntaxique des phrases dans le résumé automatique - compression de phrases narratives. *In TALN'05 : 12ème Conférence Internationale sur le Traitement Automatique du Langage Naturel*, pages 193–202, 2005.
- [YMP06] Mehdi YOUSFI-MONOD et Violaine PRINCE : Compression de phrases par élagage de leur arbre morpho-syntaxique. *Technique et Science Informatiques*, 25:437–468, 2006.