Utilisation de la Structure Morpho-Syntaxique des Phrases dans le Résumé Automatique - Compression de Phrases Narratives

Mehdi Yousfi-Monod, Violaine Prince

Utilisation de la structure morpho-syntaxique des phrases dans le résumé automatique

Compression de phrases narratives

Mehdi Yousfi Monod, Violaine Prince

LIRMM - CNRS - Université Montpellier 2 UMR 5506
161 rue Ada
34392 Montpellier Cedex 5 - France
{yousfi, prince}@lirmm.fr

RÉSUMÉ. *Nous proposons une technique de résumé automatique de textes par contraction de phrases. Notre approche se fonde sur l'étude de la fonction syntaxique et de la position dans l'arbre syntaxique des constituants des phrases. Après avoir défini la notion de constituant, et son rôle dans l'apport d'information, nous analysons la perte de contenu et de cohérence discursive que la suppression de constituants engendre. Nous orientons notre méthode de contraction vers les textes narratifs. Nous sélectionnons les constituants à supprimer avec un système de règles utilisant les arbres et variables de l'analyse morpho-syntaxique de SYGFRAN([CHA 84]). Nous obtenons des résultats satisfaisants au niveau de la phrase mais insuffisants pour un résumé complet, ce que nous expliquons en signalant la différence d'impact entre constituants et relations. Enfin nous discutons de nos travaux actuels sur un système complet de résumé, migrant depuis une notion de relation intraphrastique vers les relations interphrastiques, basé sur la suppression des paraphrases, exemples et explications, à partir des thèmes des segments textuels et sur les marqueurs lexicaux.*

ABSTRACT. *We propose an automated text summarization through sentence compression. Our approach uses constituent syntactic function and position in the sentence syntactic tree. We first define the idea of a constituent as well as its role as an information provider, before analyzing contents and discourse consistency losses caused by deleting such a constituent. We explain why our method works best with narrative texts. With a rule-based system using SYGFRAN's morpho-syntactic analysis for French ([CHA 84]), we select removable constituents. Our results are satisfactory at the sentence level but less effective at the whole text level, a situation we explain by describing the difference of impact between constituents and relations. Finally we discuss our current work dealing with a full summarization system, shifting toward a relational paradigm (intra as well as intersentential) based on paraphrases, examples and explanations deletion, by using theme information and lexical markers.*

MOTS-CLÉS : *résumé automatique, compression de phrases, analyse syntaxique*

KEYWORDS: *automatic summarization, sentence compression, syntactic analysis*

e soumission à *Technique et Science Informatiques*, le 6 avril 2005.

1. Introduction

La quantité d'informations disponibles sur Internet ou au sein de certaines entreprises, administrations et laboratoires ne cesse de croître. Ce phénomène rend la recherche d'information de plus en plus difficile. Le résumé automatique, visant à réduire considérablement la taille de ces données, apparaît comme une des solutions permettant, non seulement de faciliter cette recherche en présentant un texte pertinent de plus petite taille, mais aussi de rendre plus rapide le choix d'acceptation de la pertinence ou non d'un texte par rapport à une requête.

La production d'un résumé automatique est une tâche qui a pour principal but de le transformer en un nouveau document, plus petit, conservant les informations les plus importantes, qui sera présenté à l'utilisateur. Cette tâche pouvant s'effectuer à différents niveaux de qualité et d'objectifs, les types d'approches du résumé automatique sont multiples et variées. Elles varient tout d'abord dans le type de document à résumer (texte, image, vidéo, ...), l'application (prévisualisation, tri, rafraîchissement de mémoire, récupération de texte source, ...), la langue, le domaine, le type de public, le nombre de documents à résumer (pour un document ce sera une contraction, pour plusieurs une synthèse), la méthode de production (extraction de phrases ou de constituants[1], reformulation, ...), etc. L'ensemble des documents sources et des résumés à produire est donc très vaste et hétérogène. Ceci rend très difficile la tâche d'aborder le résumé automatique sous toutes ses facettes, simultanément. Ainsi, la majorité des approches se spécialise dans le traitement d'un type particulier de document source et la production d'un type particulier de résumé.

Les approches varient ensuite dans les techniques utilisées, mais ces dernières reflètent globalement la *démarche paradigmatique* pour laquelle deux écoles majeures de pensée s'affrontent. Une première qui prétend que l'analyse de surface est la seule efficace dans la mesure où ce sont les mots (ou les unités élémentaires) qui portent le sens, davantage que les structures. C'est la démarche majoritairement statistique. La deuxième considère que la structure des phrases est au moins aussi importante que les mots employés, et estime que la production du résumé nécessite une analyse plus ou moins profonde du document, allant de l'analyse morphologique, à l'analyse rhétorique et/ou sémantique, en passant par l'analyse syntaxique. Une nouvelle démarche, issue du fait que les documents électroniques sont de plus en plus structurés (et donc obéissent à des langages balisés) se fonde sur la récupération des balises commes indices de macrostructuration des textes permettant de les décomposer et donc de les contracter.

Dans le cas de notre étude, nous intéressons uniquement au résumé de textes bruts, et donc non balisés. De plus, nous nous proposons de lancer une première "passe" de contraction sans se préoccuper des différents types de marqueurs (lexicaux, discursifs,

1. Nous appelons *constituants* les syntagmes des phrases, c'est-à-dire toute unité de la phrase à laquelle on peut attribuer une fonction. Par exemple, prenons le groupe nominal "un médecin de famille". Il est composé de deux consituants : un groupe nominal "un médecin" et un groupe nominal prépositionnel "de famille". Ce dernier a un rôle de modificateur du premier.

sémantiques, ...), éléments que nous considérerons par la suite comme repères pour l'indication de l'importance relative d'un fragment de discours.

L'idée centrale de notre recherche est de traquer les limites de la contraction de textes par compression de phrases sans perte majeure d'information. Une fois que les avantages et les inconvénients de cette méthode auront été discutées, nous passerons à la compression de texte fondée sur des repères de plus grande granularité. En effet, il importe d'isoler les différentes variables concourant à la contraction de texte afin d'en déterminer l'impact. Si de nombreux travaux ont été réalisés dans le domaine du résumé automatique (nous en citons quelques un parmi les plus marquants ou les plus récents dans la prochaine section), peu se sont préoccupés d'évaluer le domaine de validité de la technique employée. Des travaux comparatifs existent ([LIN 03]) mais ils sont davantage orientés vers des mesures relatives sur un corpus donné, que sur une réflexion de fond sur la méthode, voire le paradigme.

La compression de phrase à partir de ses constituants n'est pas en soi une innovation. En effet, ce domaine a déjà été abordé en utilisant des méthodes statistiques ([KNI 00]), ce qui rend secondaire l'importance de la notion de composition syntaxique. En revanche, notre approche diffère fondamentalement du travail cité en ce sens qu'elle utilise la fonction syntaxique des constituants et leur position dans l'arbre syntaxique des phrases, et ne se préoccupe ni d'apprentissage, ni de fréquence relative des constituants. Sa véritable originalité est de ne pas chercher à conforter l'importance relative d'un constituant en fonction de sa fréquence, mais en fonction de son rôle syntaxique, c'est-à-dire de la relation qu'il entretient avec les autres constituants. Son importance est déterminée autant par la grammaire de la langue que par le type de texte qui le comprend.

Afin d'argumenter notre travail nous nous proposons le plan suivant. Dans la prochaine section, nous énumérons les principaux types d'approches du résumé automatique, le principe de résumé par reformulation, les résumés par extraction de phrases, puis nous comparons ceux qui travaillent à un niveau de granularité plus fin (section 2) ; nous présentons ensuite notre méthode de compression de phrases basée sur l'analyse morpho-syntaxique des phrases (section 3) ; nous continuons en illustrant l'efficacité de notre approche par les résultats d'expérimentations basées sur une application prototype appliquée à un texte du genre conte (section 4) et enfin nous terminons sur nos travaux actuels et futurs dans la réalisation d'un résumeur complet (section 5).

2. Les approches du résumé automatique

Une grande variété de techniques sont utilisées allant du résumé par extraction au résumé par reformulation. On appelle *résumé par reformulation* un texte de taille plus petite que le document auquel il se réfère, et dont le sens se veut être proche de celui du document, sans pour autant utiliser des phrases ou des portions du document initial.

La plupart des approches abordant le résumé par reformulation ([RAD 98, MCK 01, Dau 02]) sont assez semblables aux autres techniques si ce n'est qu'en fin de proces-

sus, elles transforment les informations textuelles résumées dans des formats d'entrée de modules de génération de langue comme SURGE ([ELH 96]), nous ne les détaillerons donc pas d'avantage.

Les méthodes par extraction sont fondées sur l'hypothèse "qu'il existe, dans tout texte, des *unités textuelles saillantes"* ([MIN 04]). Ces dernières représentent des points focaux, qui, soit expriment l'apport sémantique ou conceptuel du texte, soit permettent de le représenter dans sa globalité. Dès lors, le résumé par extraction cherchera à repérer ces unités saillantes et proposera un texte de taille plus petite que le document initial qui garderait majoritairement ces unités. Nous faisons également l'hypothèse de l'existence de ces unités, ainsi que de leur intérêt pour le résumé.

2.1. *Extraction de phrases clés*

La plus grande partie des approches du résumé de texte procède par extraction de phrases clés, le but étant de choisir les meilleures candidates et de les placer bout à bout pour produire le résumé final.

Une majorité de ces approches s'appuie sur des techniques statistiques, dans lesquelles des informations basées sur la fréquence des termes, comme le produit tf*idf de [SAL 73], sont fréquemment utilisées pour évaluer l'importance de chaque phrase dans un document, par exemple, les travaux relatés dans [LUH 58, BAR 97, GOL 00, BOG 00, LIN 02, RAD 04, ERK 04] l'utilisent.

Plusieurs autres techniques sont utilisées :

– les méthodes probabilistes de catégorisation, souvent assorties d'un moteur d'apprentissage (comme les Modèles de Markov), et se basant sur un corpus de documents associés à leur résumé ([JUL 95, TUR 03]) ;

– les méthodes utilisant des espaces vectoriels : par exemple [AND 00] utilise une technique proche de la décomposition en valeurs singulières (*Singular Value Decomposition, SVD*) de l'indexation sémantique latente (*Latent Semantic Indexing, LSI*) ([DEE 90]), ou encore [HIR 02] qui se base sur les *Support Vector Machines* pour séparer les phrases clés des autres ;

– les chaînes de coréférence : [BAL 98, AZZ 99] ;

– les chaînes lexicales : [CHA 01, FUE 02, ALE 03], ces approches ont tendance à prendre comme unité textuelle le paragraphe plutôt que la phrase ;

– l'utilisation de la structure rhétorique (Mann et Thompson, [MAN 88]) comme [ONO 94] qui tente de déterminer les relations rhétoriques entre les différents segments textuels (phrases, constituants) du texte source, puis conserve les noyaux des relations. La limite de cette approche est la grande difficulté à sélectionner la bonne structure rhétorique.

L'inconvénient de ces approches est que la structuration même des phrases ainsi sélectionnées n'est pas toujours compatible avec ce que l'on attend d'un résumé. En

effet, dans le cas de certains types de texte (romans, contes, ...), les phrases peuvent être longues et posséder des informations non indispensables à la compréhension du texte. Il faut donc se tourner vers d'autres types d'approches pour gérer ces cas là.

2.2. *Extraction de constituants*

Les approches précédentes utilisent des unités textuelles d'une taille au moins égale à la phrase afin de ne pas être confrontées aux problèmes d'incohérence grammaticale. Ces difficultés surviennent dans les approches dont nous allons maintenant discuter car elles travaillent à un niveau de granularité inférieur : les constituants, expressions, mots, etc. L'intérêt d'une granularité moindre est de ne pas maintenir des phrases de trop grandes d'une part, et d'autre part, de ne pas chercher à forcément supprimer une phrase donnée avant de s'assurer réellement de son aspect "superflu" par rapport au sens. Quatre orientations se trouvent principalement dans la littérature : la phrase résumé, le "copier coller", l'élagage de la structure rhéthorique (qui pourrait s'appuyer sur des textes balisés avec des langages de présentation comme XML) et la compression de phrase proprement dite.

2.2.1. *La phrase résumé*

Ce procédé consiste à extraire des segments textuels dans les différentes phrases du texte, pour former une phrase qui résume le texte.

[OKA 01] crée un graphe acyclique orienté à partir du texte source, les sommets sont des mots ou des séquences de mots et les arrêtes des relations entre les mots. Les relations se voient attribuer un score (basé sur le produit tf*idf des mots des sommets de l'arc de cette relation). Un sous-graphe est ensuite extrait, il représente la relation principale du texte. Quelques relations sont incluses au graphe afin d'ajouter quelques détails. Les mots présents dans le sous-graphe résultant sont ensuite mis bout à bout, dans le même ordre que dans le texte source, pour former la phrase résumé.

[WAN 03] se soucie du contexte dans lequel les mots extraits se trouvent afin de ne pas rassembler des mots hors-contexte. La technique utilisée se base sur la décomposition en valeurs singulières pour tenir compte de la distribution des mots et des phrases afin de regrouper les phrases touchant au même thème.

Ces techniques ne produisent que de très courts résumés (de l'ordre de la phrase) dont la cohérence grammaticale reste limitée.

2.2.2. *Le copier coller*

[JIN 00] utilise des phrases clés sélectionnées par des résumeurs classiques, les comprime (la technique sera abordée à la section 2.2.4), puis les combine en de nouvelles phrases. Les auteurs on identifié un ensemble d'opérations de combinaison à partir d'un corpus d'exemples de combinaisons réalisé par des experts, puis ont identifié un ensemble de règles de combinaison.

[ISH 02] utilise un catégoriseur SVM (Support Vector Machine) pour sélectionner les constituants à conserver pour le résumé final. Le catégoriseur est entraîné sur un corpus de phrases et un ensemble d'attributs extraits des phrases. Ces attributs sont de type genre de l'article, nombre de phrases dans l'article, position des phrases, présence des conjonctions de coordination, des démonstratifs, fréquence des termes, etc. Les constituants extraits sont ensuite rassemblés dans leur ordre original.

Les inconvénients de ces techniques sont comparables à ceux des précédentes du point de vue de la cohérence grammaticale.

2.2.3. *L'élagage de l'arbre de la structure rhétorique (SR) des phrases*

[MAR 98] utilise une combinaison d'heuristiques standard pour aider au choix de la bonne SR du texte source, au niveau inter-phrase et intra-phrase. Les sept métriques suivantes sont utilisées :

– groupement par thème : pour 2 noeuds frères de l'arbre de la SR, leurs feuilles doivent correspondre au mieux avec les frontières de changement de thèmes,

– utilisation des marqueurs : si des marqueurs sont présents dans le texte source, la SR doit les vérifier au mieux,

– groupement rhétorique par thème : identique à la première métrique si ce n'est que la comparaison se fait avec les noyaux des relations et non les feuilles,

– poids des branches situées à droite : sont préférés les arbres dont les branches droites sont plus importantes, car ce sont habituellement ces branches qui contiennent les ajouts de l'auteur moins importants et donc supprimables,

– similarité avec le titre : sont préférés les arbres dont les unités saillantes (noyaux) sont les plus similaires au titre du texte,

– position des phrases : les phrases en début ou fin de paragraphe/document sont habituellement considérées comme plus importantes ; une mesure de similarité, du même type que pour la métrique précédente, est alors effectuée,

– connexion des entités : l'information sur les relations entre les mots est prise en compte, par exemple avec les chaînes lexicales.

Selon le poids de chaque métrique utilisée dans l'heuristique, le traitement est plus efficace pour différents genres de documents. D. Marcu n'est pas parvenu à trouver une solution fonctionnant pour tout genre de texte. La cohérence est assez bien conservée dans les cas où l'analyse de la SR est correcte. La principale difficulté de cette technique reste de détecter correctement la SR.

2.2.4. *La compression de phrases*

[KNI 02] aborde le problème en utilisant un modèle de canal bruyant (*noisy-channel model*) qui consiste à faire l'hypothèse :

> (1) la phrase à comprimer fût autrefois courte et l'auteur y a ajouté des informations supplémentaires (le bruit).

Le but est alors de retrouver ces informations pour les supprimer. Les auteurs utilisent un modèle probabiliste de type modèle de Bayes qu'ils entraînent sur un corpus de documents avec leur résumé. Le moteur d'apprentissage a pour but de sélectionner les mots à conserver dans la phrase comprimée. Une faible probabilité sera attribuée à une phrase comprimée lorsque cette dernière sera incorrecte grammaticalement ou aura perdu certaines informations comme la négation. D'après leur évaluation, les résultats sont assez concluants. Relativement aux compressions réalisées par des êtres humains, une légère perte d'importance et de justesse grammaticale est observée.

[SID 02] se concentre sur la détection et la suppression des propositions relatives qui sont précédées par une proposition de la forme $GN_1\ Prep\ GN_2$, où GN_1 et GN_2 sont des groupes nominaux et $Prep$ est une préposition. Les relatives constituent, d'après Mann et Thompson, des informations sur le contexte, et ne sont donc pas indispensables. Le but est d'attacher correctement le référent de la relative en choisissant un attachement large ou local.

Par exemple, dans la phrase « *Le chien de Jean, qui a beaucoup de puces, est très joueur.* », le pronom « *qui* » se réfère à « *Le chien de Jean* » (attachement large) alors que dans la phrase « *Il est sous l'influence de Jean, qui a un fort caractère.* », le pronom « *qui* » ne se réfère qu'à « *Jean* » (attachement local).

Pour réaliser l'attachement un apprentissage machine basé sur deux probabilités d'associations est réalisé :

– entre les différentes prépositions et le type d'attachement à partir d'un corpus (Penn Wall Street Journal Treebank), par exemple la préposition *with* obtient une probabilité de 0.55 d'être en relation avec un attachement local,

– entre les deux pronoms *who* et *which* et leur référent : *who* est associé à quelque chose qui a une personnalité, contrairement à *which*. Les informations sur les relations d'hyponymes de la hiérarchie de WordNet sont utilisées.

Les résultats sur un jeu de test montrent que sa technique permet un attachement correct dans 86,2 % des cas.

Toutes ces approches basées sur des unités textuelles plus petites que les phrases ne prennent pas en compte les informations sur la fonction syntaxique et la position dans l'arbre syntaxique des constituants des phrases. Ces informations pourraient être grandement utiles dans l'aide au choix des constituants à supprimer.

[LIN 03] a évalué la qualité d'un résumé produit par extraction de phrases clés puis compression des phrases extraites. La méthode d'extraction est celle de [LIN 02] et la méthode de compression est celle de [KNI 00], basée sur le modèle de canal bruyant (similaire à [KNI 02]). Chin-Yew conclut, d'après les résultats de ses expérimentations, qu'on ne peut pas se fier à une compression strictement basée sur la syntaxe des phrases pour améliorer la qualité des résumés produits par extraction. Cependant, étant donné que Chin-Yew n'utilise que la méthode de [KNI 00] pour comprimer les phrases, nous ne sommes pas d'accord sur sa conclusion généralisée à l'ensemble

des méthodes de compression. Ce que nous concluons c'est que la méthode de compression utilisée, qui, en pratique, mélange à la fois paradigme statistique, apprentissage, technique de "noyage" (dans le bruit) et structure syntaxique, ne satisfait pas les contraintes de conservation du contenu.

Notre approche diffère grandement de celle de [KNI 00] sur au moins deux points : nos règles de compression sont produites manuellement , en relation avec des modèles linguistiques, puis mises en oeuvre, et non inférées automatiquement de façon calculatoire. De plus, nous ne faisons pas l'hypothèse de départ (1) qui pour nous est très discutable.

3. La compression de phrases par élagage de l'arbre syntaxique

Le point de départ de notre approche fût l'intuition que **la fonction syntaxique la position dans l'arbre syntaxique des constituants des phrases jouaient un rôle conséquent dans l'importance de ces constituants pour la compréhension d'un texte**. Cette intuition prend ses racines dans l'analyse grammaticale logique enseignée depuis toujours et dont on trouve des manuels connus (citons Grévisse pour mémoire). En effet, ne sont pas toujours indispensables pour comprendre le sens principal de la phrase, certains groupes adjectivaux, certains compléments circonstanciels, etc.

Par exemple, dans la phrase « *Un chat gros et laid mange une souris.* », le groupe adjectival "gros et laid" peut être supprimé sans nuire réellement à la compréhension et à l'intérêt.

Cette approche nécessite un outil d'analyse morpho-syntaxique des phrases (section 3.1) et une étude sur l'importance des constituants relativement à leur fonction syntaxique et leur position dans l'arbre syntaxique (section 3.2). Nous présentons en 3.3 l'architecture de notre système.

3.1. *L'analyseur morpho-syntaxique*

Nous utilisons l'analyseur morpho-syntaxique du français SYGFRAN, basé sur le système opérationnel SYGMART, tous deux définis dans [CHA 84]. SYGFRAN utilise un ensemble de règles de transformations d'éléments structurés, basées sur les règles de la grammaire française, qui permettent de transformer une phrase (texte brut) en un arbre syntaxique (élément structuré) enrichi d'informations sur les constituants. Cet analyseur a les avantages suivant :

– la rapidité : la complexité d'analyse est en $O(k * n * log_2(n))$ où k est le nombre de règles et n la taille du texte. Il s'agit d'une limite supérieure, car l'analyseur étant structuré en plusieurs grammaires ordonnées, le facteur multiplicatif réel est beaucoup plus petit que k. Cela dit, même ainsi, plus le texte est important, plus k est petit devant n.Aujourd'hui SYGFRAN analyse un corpus de 220000 phrases en moins d'une demi-heure.

– la robustesse : SYGFRAN parvient à obtenir une structure correcte pour au moins 30 % de l'ensemble des différents cas de syntaxe des phrases du français, pour les autres cas, **SYGFRAN fournit une analyse partielle mais exploitable**.

– la production d'un arbre syntaxique : la plupart des systèmes actuels d'analyse syntaxique ne réalisent qu'un simple marquage linéaire, ceux qui produisent un arbre n'ont qu'une très faible couverture sur l'ensemble des constructions syntaxiques existantes.

SYGFRAN prend en entrée du texte brut et produit une structure parenthésée, correspondant à l'arbre morpho-syntaxique de chaque phrase du texte, dans laquelle de nombreuses variables sont renseignées sur les différents natures, fonctions syntaxiques, formes canoniques, catégories grammaticales, temps, modes, genres, nombres, etc. des constituants.

Par exemple, l'analyse de cette phrase produit l'arbre syntaxique de la figure 1.

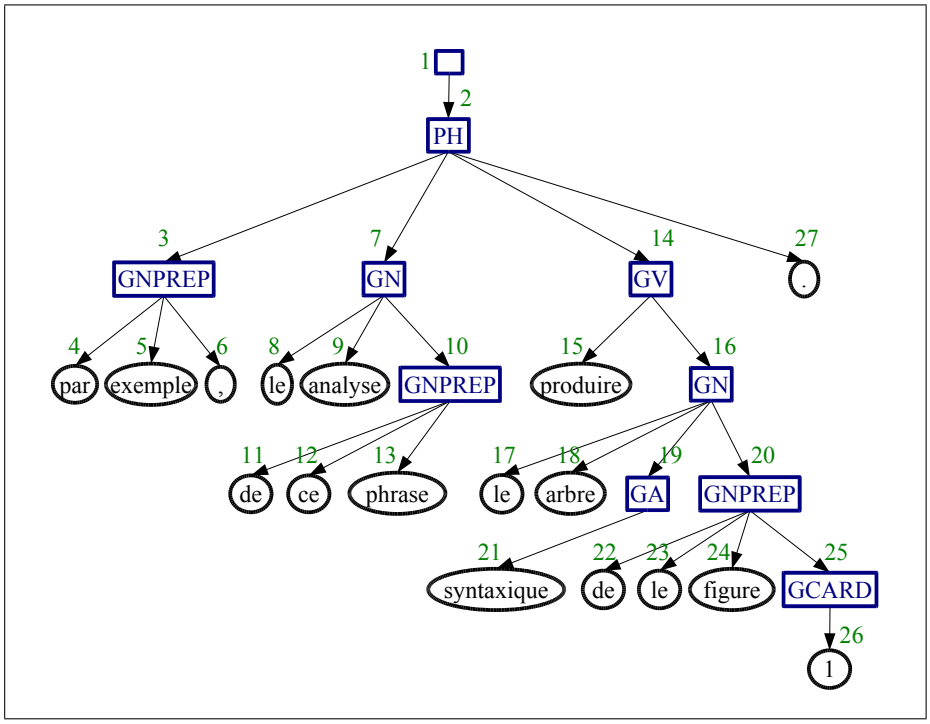

Figure 1. *Exemple d'analyse de SYGFRA*

Les noms des noeuds internes (rectangles) correspondent aux natures des **constituants** : PH pour PHrase, GN pour Groupe Nominal, GV pour Groupe Verbal, GA pour Groupe Adjectival, $GNPREP$ pour Groupe Nominal PRÉPositionnel et $GCARD$ pour Groupe CARDinal. Les noms des feuilles (ellipses) sont les formes

canoniques des lexies (masculin, singulier, infinitif). Le numéro de chaque noeud est un pointeur sur les informations des variables SYGFRAN associées au noeud.

Par exemple, le noeud 3 possède entre autres les variables et valeurs suivantes :

Variable	**Valeur**	**Description**
GNR	MAS	le genre est "masculin"
NUM	SIN	le nombre est "singulier"
CAT	N	la catégorie (des éléments simples) est "nom"
K	GNPREP	la catégorie (des groupes) est "groupe nominal prépositionnel"
FS	COMPCIR	la fonction syntaxique est "complément circonstanciel"

Le noeud numéro 1 est le père des phrases du document, dans notre cas il n'y a qu'une phrase.

Lorsque SYGFRAN ne parvient pas à produire l'intégralité de la structure syntaxique d'une phrase ou d'un constituant c, il crée un noeud de nom "ULFRA", qui signifie unité linguistique française de nature indéterminée, auquel il ajoute, pour chaque sous-constituant s de c, l'arbre syntaxique de s.

La figure 2, basée sur cette phrase, illustre ce cas.

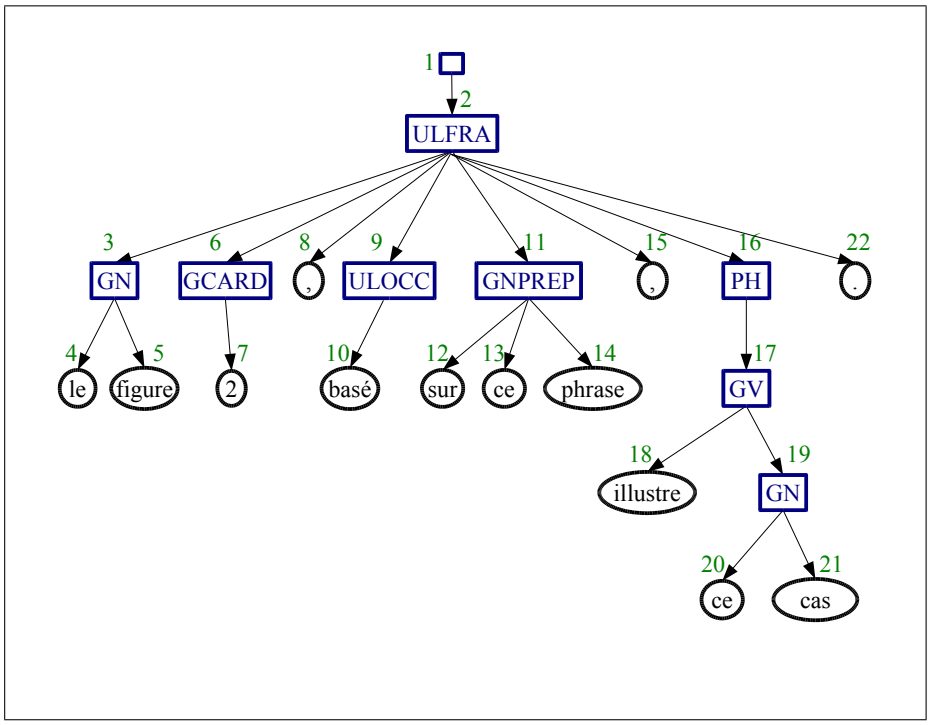

Figure 2. *Exemple d'analyse ayant partiellement échoué*

Dans cet exemple, c'est la locution "être basé sur" que l'analyseur ne connaît pas et n'arrive donc pas à analyser correctement. Il spécifie "ULOCC" (*Unknown Locution*) comme nom du noeud père du mot "basé" pour exprimer son incompréhension. La structure de certains constituants reste tout de même correcte et peut donc être exploitée. Par exemple, l'arbre ayant pour racine le noeud 11, contient l'information que "sur cette phrase" est un groupe prépositionnel.

3.2. *Fonction et Position*

Le test de suppression des constituants est abordé par de nombreux ouvrages sur la grammaire française pour aider à la détermination de la fonction syntaxique d'un constituant. Le test est validé si la phrase résultante reste grammaticalement cohérente. Cependant, les textes linguistiques traitant de l'importance des constituants dans la phrase selon leur fonction syntaxique sont beaucoup plus rares. Des recommandations sont fournies par les linguistes, mais pas de règle fondamentale. Nous avons donc procédé de la manière suivante. Nous avons considéré ces recommandations comme des hypothèses de travail et nous avons cherché à les étayer empiriquement. Ainsi, Melçuk, dans son analyse du Français contemporain, parle de fonctions syntaxiques dite de "gouvernement" (à la suite des travaux de Chomsky). Sont **gouverneurs** des constituants considérés comme indispensables à la cohérence grammaticale et sémantique de la phrase. Ainsi, le sujet d'une phrase et son groupe verbal sont gouverneurs sur le plan de la cohérence grammaticale. Considérons la phrase simple suivante :

Jean mange une pomme verte.

Le sujet "Jean", s'il est supprimé, produit une phrase incohérente. Comme il est atomique, on ne peut pas le réduire. Le verbe "mange" également. Si on supprime le complément d'objet direct, "une pomme verte", on a une phrase grammaticalement cohérente (car le verbe *manger* a une forme intransitive) en revanche, on perd de l'information importante, vu que le verbe n'est pas utilisé ici de manière intransitive. Il est spécifiquement qualifié, il importe donc de lui restituer son complément, sur lequel on regarde si on peut appliquer une fonction de restriction.

Dans le constituant "une pomme verte" il y a en réalité deux constituants, qui se divisent à leur tour en gouverneur et non gouverneur. Dans un groupe nominal adjectival, le nom est gouverneur et la restriction "une pomme" par rapport à "une pomme verte" ne perd pas en cohérence grammaticale et ne perd pas sa fonction syntaxique.

Ainsi la détermination du constituant *secondaire* se fait par rapport au rôle syntaxique. Trois niveaux de granularité sont considérés, la **phrase** (qui peut comprendre plusieurs propositions), la **proposition** (qui est définie par un sujet, un verbe et éventuellement un ou plusieurs compléments) et le **constituant nomina**l.

Au niveau de la phrase, l'importance d'un élément est attribuée selon l'ordre suivant :

– La proposition principale :

Jean mange une pomme verte.

– les propositions relatives tenant lieu de complément du verbe :

Jean mangera une pomme verte quand la saison des pommes arrivera.

– les propositions relatives tenant lieu d'épithète, et se trouvant généralement en apposition (entre deux virgules, juste après le nom qu'elles sont censées qualifier) :

Jean, qui attendait l'arrivée de son frère, mangeait une pomme verte.

Sont considérés, dans l'ordre d'importance, en tant que relations, au sein d'une proposition :

– les sujets et verbes,

– les compléments d'objet (directs et indirects),

– les compléments circonstanciels.

A l'intérieur même d'un constituant nominal, sont considérés, dans l'ordre d'importance :

– les noms ,

– les compléments de noms,

– les adjectifs (épithètes).

L'idée est de dire que plus on descend dans la liste (par rapport à une granularité donnée) plus on a de chances de réaliser une compression sans perte de cohérence ni perte d'information. Tout le problème consiste à savoir si :

– on peut supprimer systématiquement ou non des élément de granularité plus large comme les propositions relatives,

– on peut supprimer les moins importants des constituants (les compléments circonstanciels par exemple),

– on peut élaguer des constituants nominaux,

si ces actions peuvent être relativement généralisées (grosso modo, à tout type de texte).

Pour cela, à partir de textes de genres variés, nous avons réalisé des tests de suppression de certains constituants en fonction de leur fonction syntaxique (donc plutot la granularité "moyenne"), en estimant les pertes de cohérence discursive et de contenu important dans les phrases comprimées.

Dans les textes du genre article scientifique ou énoncé technique, chaque constituant se révèle avoir beaucoup plus d'importance que dans un texte narratif (roman, conte, ...). Par exemple, si on prend le terme « *hormone de synthèse* », il serait très ennuyeux de supprimer le complément de nom. De la même manière, il serait gênant d'amputer la phrase « *Un vent de 50 kmh soufflera sur le Golfe du Lion.* » de son complément circonstanciel de lieu ("le Golfe du Lion"). En revanche, dans « *L'étalon noir*

broutait, tranquillement, en remuant la queue, près de l'enclos principal. », il est tout à fait possible de réduire cette phrase sans perte d'information risquant d'en transformer le sens. La raison est que les auteurs de textes narratifs ajoutent de nombreuses informations à caractère essentiellement descriptif qui aident le lecteur à être transporté dans l'histoire mais qui ne sont pas indispensables à la compréhension du coeur de l'histoire. Alors que dans un article scientifique ou technique, chaque constituant a un rôle important à jouer dans la compréhension du discours. Afin d'évaluer les qualités de la compression par suppression de constituants, nous avons donc cherché à la tester sur des corpus où elle avait un sens, en d'autre termes dans les textes de type **narratif**, en se proposant ultérieurement de tester d'autres paradigmes pour les textes scientifiques ou techniques.

[MAN 04] aborde la problématique du résumé de textes narratifs, en s'appuyant principalement sur des indices temporels. Il étudie les évènements sur trois plans : la scène, l'histoire et l'intrigue, dans le but d'extraire les évènements clés, scènes clés, et les intrigues saillantes. Il compte sur les méthodes actuelles (basées sur le marquage lexical, l'étude de la structure rhétorique, l'analyse morpho-syntaxique, ...) et futures pour extraire les indices temporels nécessaires. Notre méthode actuelle ne tient compte que des informations syntaxiques.

Les deux facteurs (cohérence et importance) varient selon le genre de texte et le type des constituants. En supprimant dans une première passe les constituants les plus secondaires on obtient un résumé dont le contenu important est bien conservé mais dont la taille est grande. La compression peut alors consister à plusieurs passes jusqu'à obtenir un rapport spécifique (taille/pertes) du résumé produit. Chaque constituant est supprimé par élagage de l'arbre syntaxique. Après une première passe, les arbres syntaxiques obtenus se révèlent être de bons représentants des originaux. Leur représentativité se dégrade sensiblement après chaque passe.

Nous avons noté trois catégories de constituants susceptibles d'être supprimés selon leur fonction syntaxique et leur position : les compléments circonstanciels (section 3.2.1), les épithètes (section 3.2.2) et les appositions (section 3.2.3). Comme on peut le voir, ils sont de granularité moyenne. Les appositions, lorsqu'elles se transforment en propositions relatives (complément de nom) deviennent de granularité plus importante, et augmentent de ce fait le taux de compression obtenu.

3.2.1. *Les compléments circonstanciels*

L'importance des différents compléments circonstanciels (CC) dépend du type de texte. De manière générale, ce sont les CC de *temps* et de *but* qui se sont révélés être les plus importants. La raison est qu'ils répondent aux questions que nous jugeons les plus importantes à savoir "Quand ?" et "Dans quel but ?". Les CC de *lieu* (questions "Où ?") ont leur importance principalement au début du texte, lorsque le décor est posé. Ceux de *manière* (questions "Comment ?") et de *cause* (questions "Comment est-ce arrivé ?") sont peu importants dans une majorité des cas. La fréquence d'apparition des autres CC (*comparaison, condition, conséquence, opposition, mesure, ...*) étant assez faible, leur suppression n'aboutit fréquemment qu'à une petite perte de contenu.

Certains gérondifs fonctionnent comme des propositions subordonnées circonstancielles, nous les supprimons aussi. Exemple : « *Jean mange des bonbons en chantant.* »

L'importance des CC varie aussi selon la nature du verbe de la proposition. Dans le cas d'un CC de lieu placé après le verbe "être", la suppression ne sera pas possible. Par exemple, on ne peut supprimer le CC de la phrase « *Jean est dans la voiture.* ». Cependant, si plusieurs CC de lieu se suivent après le verbe être, tous sauf un pourront être supprimés sans grande perte de contenu : « *Jean est dans la voiture, dans le garage de sa maison, près de la confiserie.* »

Enfin nous avons remarqué qu'un CC situé dans une phrase interrogative était très important car la question porte généralement sur lui. Exemple : « *Jean n'aurait-il pas dormi dans la confiserie ?* »

3.2.2. *Les épithètes*

Les adjectifs et groupes adjectivaux, mais aussi certaines propositions relatives (complément de nom), ont une fonction d'épithète. Par exemple, dans la phrase « *L'enfant qui mange des bonbons paraît heureux.* », le constituant souligné est une relative qui a une fonction d'épithète. D'une manière comparable aux CC, lorsqu'un épithète est placé après le verbe "être", et plus généralement après un verbe d'état, son importance s'accroît considérablement, rendant la suppression impossible. Enfin, nous avons noté que lorsque l'épithète était placé dans un groupe nominal dans lequel le déterminant était un article défini, alors sa suppression était difficile. Ceci est dû au fait que l'article défini est utilisé pour parler d'une entité particulière et que les épithètes du nom permettaient de différencier cette entité des autres.

Par exemple, dans la phrase « *Il y avait deux enfants devant moi, l'enfant blond s'est approché.* », l'adjectif épithète souligné permet de préciser quel enfant s'est approché. Supprimer cet adjectif causera une perte de contenu considérable. En revanche, dans la phrase « *Il vit un enfant blond dans la rue.* », l'adjectif est moins important et peut donc être supprimé.

3.2.3. *Les appositions*

L'apposition peut avoir des natures variées, elle peut être :

– un groupe nominal (« *Jean, le gourmand, aime les bonbons.* »),

– un pronom (« *Jean doit manger lui-même les bonbons.* »),

– une proposition relative (cf. section 3.2.2),

– une proposition participale présent (« *Jean, aimant les bonbons, a beaucoup de caries.* »),

– une proposition participale passé (« *Jean, aimé des enfants, fera un bon père.* »),

– une proposition infinitive (« *Jean, manger des légumes, cela m'étonnerait !* »).

Dans les trois premiers cas, les constituants se suppriment sans difficulté. Les propositions participales sont aussi sujettes à la suppression, mais une perte un peu plus

importante de contenu est à noter. Dans le dernier cas, la suppression paraît difficile car la proposition infinitive apporte systématiquement un information importante qui vient compléter le sujet.

3.3. *Architecture*

L'architecture de notre système est présenté en figure 3.

Du texte source sont produits les arbres syntaxiques correspondant au résultat de l'analyse faite par SYGFRAN. Ensuite, le module de sélection/coloration de segments textuels utilise les informations suivantes pour effectuer la sélection :

– le texte source,

– les arbres syntaxiques et les variables/valeurs fournis par SYGFRAN

– le seuil du rapport taille/pertes à ne pas dépasser fourni par l'utilisateur ou défini par le type d'application,

– l'ensemble des règles de sélection des constituants,

pour effectuer les différentes passes de sélection des constituants jusqu'à satisfaction du rapport taille/pertes. Les constituants sélectionnés sont ensuite supprimés.

4. Expérimentations

Nous avons réalisé un programme prototype afin de pouvoir mesurer l'efficacité d'une telle approche. Nous avons défini un système utilisant des règles simples, basées sur les résultats de notre étude expérimentale (section 3.2). Chaque règle possède un nom auquel on associe un ensemble de couples (clé,valeur). Chaque nom représente un type de constituant susceptible d'être supprimé. Les couples (clé,valeur) sont les contraintes qu'un constituant doit respecter pour être sélectionné à la suppression. Notre système actuel possède trois types de contraintes :

– une sur la valeur de la variable du constituant fournie par SYGFRAN (par exemple, le constituant doit être un complément circonstanciel),

– une sur la position du constituant par rapport à un autre constituant relativement à un noeud père spécifique (par exemple, le constituant ne doit pas être à droite d'un verbe d'état),

– une sur la position du constituant par rapport à un antécédent possédant une valeur spécifique à une clé (par exemple, le constituant ne doit pas être un sous-constituant d'une phrase interrogative).

Notre prototype actuel n'effectue qu'une passe. Nous comptons créer par la suite des règles paramétrables afin de gérer plusieurs rapports de taille/pertes dans la production du résumé.

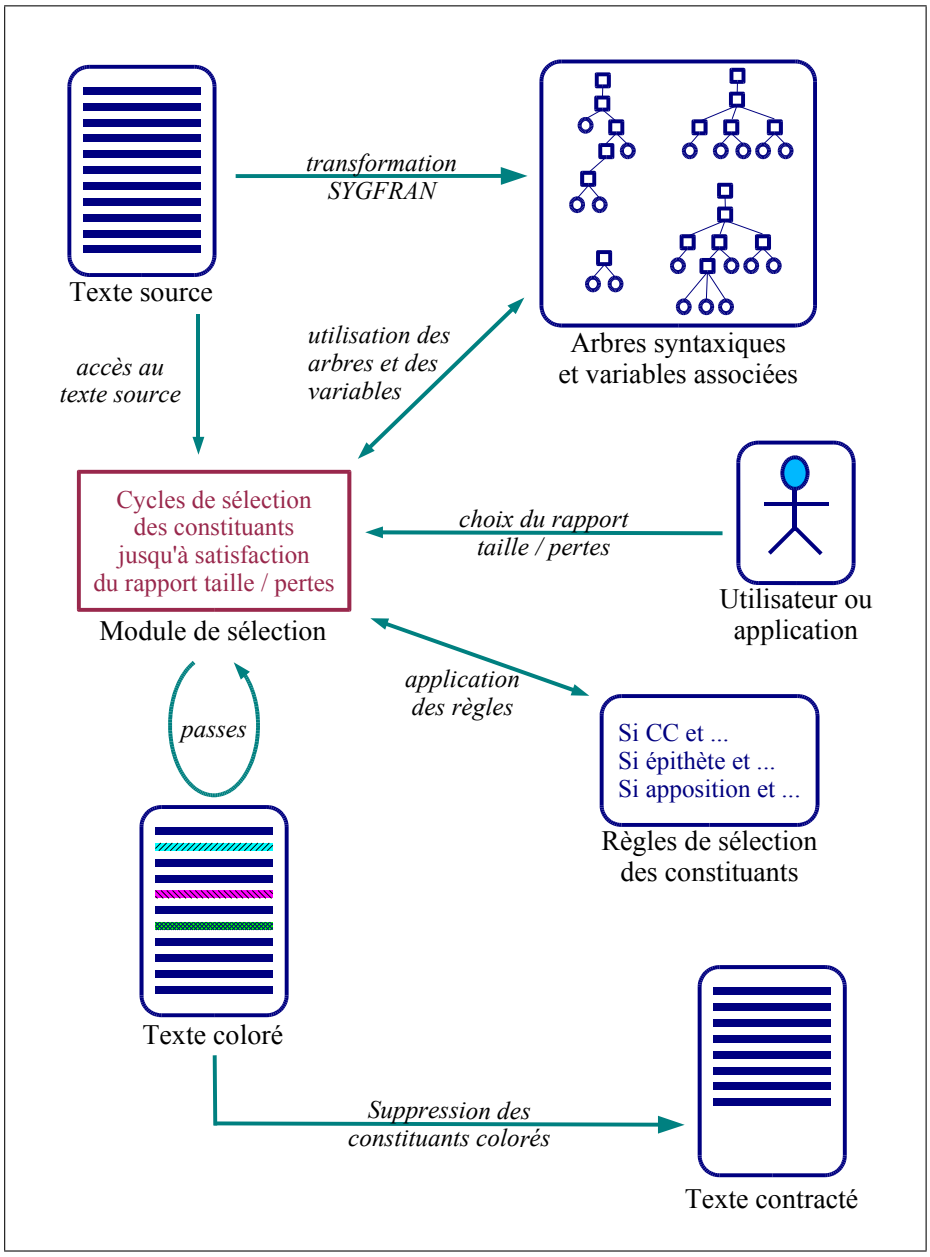

Figure 3. *Fonctionnement de notre système de compression de phrases*

La première phase consiste à colorier les constituants susceptibles d'être ôtés par la suite. Une couleur est attribuée à chaque type de constituant. Ainsi il est aisé d'estimer la qualité des règles sur le texte en cours avant de supprimer réellement ces constituants. Dans la seconde phase, les segments textuels colorés sont supprimés pour obtenir le résumé final.

Nous avons créé un jeu de test de règles (figure 4(a)), utilisant les variables décrites dans la figure 4(b) et les valeurs décrites dans la figure 4(c).

La première règle peut être traduite par : je nomme et sélectionne $compcir$ (CC) tout constituant qui :

– a "complément circonstanciel" pour fonction syntaxique (c'est-à-dire qui est un CC),

– n'a pas "temps" pour sémantique de l'objet (c'est-à-dire qui n'est pas un CC de temps),

– n'a pas un antécédent de type phrase interrogative (c'est-à-dire qui n'est pas inclus dans une phrase interrogative) et

– vérifie soit :

- n'est pas situé à droite d'un constituant qui a "verbe d'état" comme interprétation des constructions syntaxiques, relativement à un noeud père qui a "groupe phrase" pour catégorie des groupes (c'est-à-dire qui ne soit pas précédé d'un verbe d'état)

ou

- est situé à droite d'un constituant qui a "complément circonstanciel" comme fonction syntaxique, relativement à un noeud père qui a "groupe phrase" pour catégorie des groupes (c'est-à-dire qui soit à droite d'un autre CC).

Dans la première phrase du texte de la figure 5, le constituant "même s'il faisait des bêtises", souligné d'un trait simple, a été sélectionné par cette règle car il vérifie l'intégralité de ses contraintes : c'est un complément circonstanciel qui n'est pas de temps, qui n'est pas dans une phrase interrogative et qui n'est pas situé à droite d'un verbe d'état.

Nous avons utilisé comme texte de test un conte haïtien. La principale raison de ce choix est que SYGFRAN produit une syntaxe correcte pour l'intégralité des phrases de ce texte. Le résultat de la coloration de la première moitié de ce texte est présenté en figure 5.

4.1. *Discussion sur les résultats*

Avec le jeu de règles actuel, notre approche nous a permis d'éliminer environ 34 pourcent du texte complet. Ce résultat est déjà très intéressant mais rarement suffisant en termes de taille du texte résumé.

Nous constatons une légère perte de contenu et de cohérence discursive, celle-ci reste plus que raisonnable au regard des techniques actuelles de résumé automatique. La cohérence grammaticale, quand à elle, est très bien conservée. Nous estimons que les règles peuvent encore être affinées, mais les données linguistiques dans ce domaine sont très limitées.

Pour ce texte, SYGFRAN nous fournit des arbres syntaxiques corrects, mais les valeurs des variables ne sont pas systématiquement justes et complètes. Pour les CC, SYGFRAN ne spécifie actuellement la sémantique de l'objet que pour ceux de temps et de lieu.

Pour le constituant "afin de découvrir où elle allait" du deuxième paragraphe, nous possédons l'information que c'est un CC mais pas que c'est un CC de but. Ce genre de constituant devrait être conservé. Dans le cas du constituant "D'habitude" du troisième paragraphe, SYGFRAN ne détecte pas que c'est un CC de temps, c'est pourquoi nous le sélectionnons à tort à la suppression. Idem pour "Finalement" au quatrième paragraphe.

L'évolution des règles de SYGFRAN permettra de gérer de tels cas.

5. Travaux actuels et futurs

Les règles de sélection des constituants à supprimer peuvent être affinées davantage selon la fonction des constituants et surtout selon le genre des textes. Nous comptons, à cet effet, effectuer des expérimentations sur plus de textes touchant à des genres plus variés.

Cependant, la compression de phrases ne suffit pas à produire un résumé d'une taille convenable dans la plupart des cas d'applications. Comme nous l'avons vu, elle est aussi fortement dépendante du genre de texte. Si l'objectif du résumé est de rendre un peu plus concis un texte trop bavard, ou, pour des raisons d'espace pour l'affichage, d'obtenir un texte un peu plus court, alors cette compression peut suffire. Mais ces cas particuliers sont rares, et, dans la majorité des cas, une contraction plus efficace en taille sera nécessaire.

Dans cette optique, nos travaux actuels portent sur le résumé par suppression de segments textuels basée sur leur fonction dans le discours. Nous estimons que les fonctions discursives utiles au résumé incluent celles d'exemplification, de paraphrasage et d'explication car ces parties ont pour but de faciliter la compréhension du lecteur et leur suppression ne causera donc pas des pertes importantes. La taille de ces segments textuels peut varier du constituant à un ensemble de phrases. Pour détecter ces segments et leur limites textuelles, nous comptons utiliser conjointement deux informations : le thème des segments (section 5.1) et les marqueurs lexicaux (section 5.2).

Nous souhaitons aussi utiliser la compression de phrases comme pré ou post traitement à cette nouvelle approche. La question est de savoir si la compression de phrases

dégradera ou améliorera les performances des techniques basées sur le thème ou sur les marqueurs lexicaux. Dans le cas où il y a dégradation, il faudra utiliser la compression de phrases en post-traitement. Dans le cas contraire, ce sera en pré-traitement.

5.1. *Le thème des segments textuels*

Nous pensons que si un segment textuel est le paraphrasage ou l'explication d'un autre, alors les deux segments ont un thème très proche. Par **thème** nous désignons la représentation du sens des phrases composant le segment textuel, représentation qui se fonde sur un calcul à partir des sens des mots qui appartiennent à ces phrases ainsi que les contraintes fournies par la structure même du segment. Ces dernières sont souvent appelées contraintes de désambiguisation. Elles agissent comme une aide à l'interprétation, puisque les mots peuvent avoir plusieurs sens possibles (polysèmes) et c'est leur mise en commun ainsi que la structure de la phrase qui permettent de pencher vers un sens *préférentiel*, aussi appelé sens en contexte.

Les vecteurs sémantiques de [CHA 03] permettent de donner une information sur le thème d'un segment textuel, nous comptons donc les utiliser dans cette tâche de localisation. Sans pour autant rentrer dans une description longue de cette méthode de représentation du sens, voici quelques définitions qui permettent de donner une idée générale de ce que nous réalisons.

Chaque mot du dictionnaire de la langue est représenté à l'aide d'un vecteur de valeurs booléennes plongé dans un espace à 873 dimensions, qui correspondent au 873 concepts fondateurs du thésaurus Larousse. Ces valeurs, quand elles sont non nulles, traduisent le fait que le concept concerné participe à l'élaboration du sens du mot. Par exemple, le terme *voile* est indexé par les concepts de VIE DOMESTIQUE (pour *voile de rideau*) de TRANSPORT (pour *voile de bateau*) d'EMOTION (pour *voile de deuil*) et de RELIGION (pour *porter le voile*). La structure arborescente de la phrase, déterminée par SYGFRAN, ainsi que les différentes relations qu'entretiennent les groupes (gouverneurs ou non) permettent de fournir une combinaison linéaire pondérée des différents vecteurs de mots qui vont représenter une première description vectorielle de la phrase. Ce vecteur de phrase est ensuite utilisé comme un renforçateur pour déterminer le sens en contexte des mots, et de nouveau un vecteur de phrase contextualisé est calculé.

Le vecteur de segment est soit un vecteur de phrase, soit le vecteur **centroïde** des phrases du segment. On peut alors facilement calculer une distance entre vecteurs (comme cela a été présenté dans [CHA 03] pour la catégorisation de documents). En toute logique, il n'est pas absurde de considérer que si cette distance est faible, alors la représentation thématique, issue du calcul sémantique vectoriel des deux segments, matérialisés par leurs vecteurs, est elle aussi très proche. On dira, sans grand danger d'erreur, que les deux segments "parlent de la même chose".

En revanche, l'information du thème s'applique plus difficilement aux exemples car on peut illustrer une situation, un principe par des exemples de thématiques très va-

riées. C'est pourquoi la proximité, voire l'identité thématiques ne sont pas suffisantes pour éliminer des segments. Il faut alors se concentrer sur des repères permettant de définir l'importance relative des segments, à l'instar de la notion de constituant secondaire que nous avons relevés dans la structure intra-phrastique.

5.2. *Les marqueurs lexicaux*

Dans les textes, en particuliers scientifiques, ce qui peut jouer le rôle de segment incident, ou secondaire, ce sont les exemples. Nous avons choisi, pour localiser les exemples, de nous aider de marqueurs lexicaux : "par exemple","comme", "e.g" sont les marqueurs les plus communs. Dans la mesure où les ouvrages linguistiques abordant ce sujet sont plutôt rares, il n'est donc pas impossible que nous analysions nous-mêmes les différents cas possibles. En pratique, nous avons trouvé une référence, en psycholinguistique ; il s'agit de [DEL 97], qui réalise une ébauche d'étude linguistique des différents marqueurs et formes des exemples. Nous chercherons donc nous appuyer sur ce texte pour modéliser notre approche.

De la même manière, d'autres segments, dont les paraphrases ou explications, peuvent être marqués. Les marqueurs les plus connus sont des locutions telles que : "en d'autres termes", "c'est-à-dire", "autrement dit" ou des locutions latines comme "id est". Pour les explications, outre les marqueurs de paraphrase, il y a les marqueurs argumentatifs comme : "car", "c'est pourquoi", "dès lors", etc.

Cela dit, les marqueurs souffrent d'un certain nombre de limites. Tout d'abord ils ne fournissent qu'un indice de présence éventuelle d'un segment de type exemple ou paraphrase ; la principale difficulté sera de localiser précisément le début et la fin du segment textuel correspondant. Deuxièmement, ils peuvent être ambigus : "comme" désigne un exemple ou un complément circonstanciel ou une énumération comparative, ce qui revient au problème du repérage du segment incident (phrase ou paragraphe). Enfin les marqueurs ne sont pas seulement lexicaux, ils peuvent être lexicographiques : les parenthèses, incises, virgules, et points virgules peuvent jouer des rôles non négligeables pour déterminer l'importance relative des segments. Comment alors procéder pour automatiser le traitement du marquage ? La question reste encore ouverte.

Disons que, en règle générale, les marqueurs sont une piste à ne pas négliger, mais il est clair qu'ils n'ont pas une fiabilité maximale. Notre but n'est pas de faire une théorie automatisée du marquage, mais d'utiliser au mieux ce dernier pour la contraction de textes. La détection du marqueur devra être confortée par d'autres heuristiques. Par exemple, dans un article scientifique, nous nous intéresserons beaucoup aux marqueurs qui ont en plus un rôle dans la présentation du document (sous forme de sous-titre, ou par une exergue en gras).

6. Conclusion

Bien que le problème du résumé automatique ait déjà été abordé par de nombreux scientifiques depuis presque 50 ans [LUH 58], l'approche que nous avons adoptée est novatrice. Les approches actuelles du résumé automatique utilisent des informations telles la fréquence des termes, les relations lexicales entre les termes, les étiquettes sur la nature des constituants fournis par des *POS tagger* (lemmatiseurs), les probabilités d'un constituant d'apparaître dans un résumé d'après des moteurs d'apprentissage, la structure rhétorique du texte, cependant, aucune d'entre elles n'utilise conjointement **la fonction syntaxique et la position dans l'arbre syntaxique des constituants**.

Ces informations n'ont pas été réellement exploitées jusqu'à présent car elle ne peuvent être extraites qu'avec des analyseurs morpho-syntaxiques fonctionnant avec un niveau suffisant. Ce niveau n'a été atteint que relativement récemment en traitement automatique des langues, parce qu'il est fort coûteux en temps de calcul. L'amélioration drastique de la technologie des processeurs et leur rapidité de traitement a permis l'émergence d'analyseurs de qualité suffisante pour aborder le problème de l'analyse en constituants. Le système opérationnel SYGMART est l'un de ces outils. En outre, il ajoute à l'analyse en constituants de nombreuses informations concernant les relations entre constituants, ce que peu d'autres analyseurs proposent.

Notre approche a débuté par une étude sur l'importance des constituants dans une phrase. Le critère de suppression a été l'évaluation de la perte de contenu et de cohérence que la suppression de ces constituants engendre. Le critère de sélection est celui de la fonction syntaxique et de la position dans l'arbre syntaxique des constituants. Les textes narratifs (romans, contes, ...) se sont révélés être les plus adéquats pour une telle approche.

Nous avons alors modélisé une compression de phrases basée sur la suppression de ces constituants. La création d'un système de règles basé sur notre modélisation nous a permis de tester la faisabilité d'une telle approche. Nous sommes passés par une étape de coloration des constituants en fonction des règles qui les avaient sélectionnés, afin d'estimer la pertinence de chaque règle. Notre méthode nous a permis de supprimer environ 34 pourcent du texte de test, tout en conservant une très bonne cohérence grammaticale.

Nous avons conclu que notre compression est utile dans des cas d'application précis et peut être utilisée dans un processus plus large de résumé automatique. Celui vers lequel nous nous orientons a été abordé, il se base sur la suppression de segments textuels de type paraphrases, exemples et explications. Nous comptons détecter de tels segments en utilisant des vecteurs sémantiques et des marqueurs lexicaux. Les techniques qui seront employées deviennent suffisamment éprouvées, car elles sont en permanence évaluées et améliorées dans d'autres sous-domaines du traitement automatique des langues, comme en catégorisation de documents par analyse de contenu. En rendant au résumé automatique son rôle d'application du traitement automatique des langues nous espérons obtenir des résultats plus adaptés aux besoins des lecteurs

humains des textes, et d'une qualité plus sûre que celle fournie par les démarches purement quantitatives et fréquentielles.

7. Bibliographie

[ALE 03] ALEMANY L. A., FORT M. F., « Integrating Cohesion and Coherence for Automatic Summarization », *EACL03*, Budapest, Hungary, avril 2003.

[AND 00] ANDO R., BOGURAEV B., BYRD R., NEFF M., « Multi-document summarization by visualizing topical content », *in Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization*, 2000.

[AZZ 99] AZZAM S., HUMPHREYS K., GAIZAUSKAS R., « Using coreference chains for text summarization », *in Proceedings of the ACL'99 Workshop on Coreference and its Applications*, Baltimore, 1999.

[BAL 98] BALDWIN B., MORTON T., « Dynamic coreference-based summarization », *in Proceedings of EMNLP-3 Conference*, 1998.

[BAR 97] BARZILAY R., ELHADAD M., « Using lexical chains for text summarization », *in Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, Madrid, Spain, 1997, ACL.

[BOG 00] BOGURAEV B. K., NEFF M. S., « Lexical Cohesion, Discourse Segmentation and Document Summarization », *RIAO-2000*, Paris, avril 2000.

[CHA 84] CHAUCHÉ J., « Un outil multidimensionnel de l'analyse du discours », *in Coling'84*, Standford University, California, 1984, p. 11-15.

[CHA 01] CHAVES R. P., « WordNet and Automated Text Summarization », *in Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS*, Tokyo, Japan, 2001.

[CHA 03] CHAUCHE J., PRINCE V., JAILLET S., TEISSEIRE M., « Classification automatique de textes à partir de leur analyse syntaxico-sémantique », *in Proceedings of TALN'2003*, vol. 1, Batz-sur-mer, 2003, p. 45-55.

[Dau 02] DAUMÉ III H., ECHIHABI A., MARCU D., MUNTEANU D. S., SORICU R., « GLEANS : A Generator of Logical Extracts and Abstracts for Nice Summaries », *in Proceedings of the Document Understanding Conference (DUC-2002)*, Philadelphia, PA, juillet 2002.

[DEE 90] DEERWESTER S. C., DUMAIS S. T., LANDAUER T. K., FURNAS G. W., HARSHMAN R. A., « Indexing by Latent Semantic Analysis », *Journal of the American Society of Information Science*, vol. 41, n° 6, 1990, p. 391-407.

[DEL 97] DELCAMBRE I., *L'exemplification dans les dissertations*, Presses Universitaires Du Septentrion, Villeneuve d'Ascq, 1997.

[ELH 96] ELHADAD M., ROBIN J., « An overview of SURGE : a re-usable comprehensive syntactic realization component », *in Proceedings of the 8th International Workshop on Natural Language generation (demonstration session) (INLG'96)*, Brighton, UK, 1996.

[ERK 04] ERKAN G., RADEV D. R., « LexRank : Graph-based Centrality as Salience in Text Summarization », *Journal of Artificial Intelligence Research (JAIR)*, , 2004.

[FUE 02] FUENTES M., RODRÍGUEZ H., « Using cohesive properties of text for Automatic Summarization », *in Proceedings of the Primeras Jornadas de Tratamiento y Recuperación*

de Información (JOTRI2002), Valencia, Spain, 2002.

[GOL 00] GOLDSTEIN J., MITTAL V., CARBONELL J., KANTROWITZ M., « Multi-document summarization by sentence extraction », *in Hahn et al.[15]*, 2000, p. 40-48.

[HIR 02] HIRAO T., ISOZAKI H., MAEDA E., MATSUMOTO Y., « Extracting Important Sentences with Support Vector Machines », *in Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, août 2002, p. 342-348.

[ISH 02] ISHIKAWA K., ICHI ANDO S., ICHI DOI S., OKUMURA A., « Trainable Automatic Text Summarization Using Segmentation of Sentence », *in Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering*, 2002.

[JIN 00] JING H., MCKEOWN K., « Cut and paste based text summarization », *in Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000, p. 178-185.

[JUL 95] JULIAN K., O. P. J., FRANCINE C., « A Trainable Document Summarizer », *in Proceedings of the 18 th ACM SIGIR conference on research and development in information retrieval*, 1995, p. 68-73.

[KNI 00] KNIGHT K., MARCU D., « Statistics-Based Summarization - Step One : Sentence Compression », *in Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, Sapporo, Japan, 2000, p. 703-710.

[KNI 02] KNIGHT K., MARCU D., « Summarization beyond sentence extraction : a probabilistic approach to sentence compression », *Artificial Intelligence archive*, vol. 139(1), 2002, p. 91-107.

[LIN 02] LIN C.-Y., HOVY E. H., « Automated Multi-Document Summarization in NeATS », *in Proceedings of the DARPA Human Language Technology Conference*, 2002, p. 50-53.

[LIN 03] LIN C.-Y., « Improving Summarization Performance by Sentence Compression - A Pilot Study », *in Proceedings of the Sixth International Workshop on Information Retrivial with Asian Language (IRAL 2003)*, Sapporo, Japan, juillet 2003.

[LUH 58] LUHN H., « The automatic creation of literature abstracts. », Journal of research and development, 1958, IBM.

[MAN 88] MANN W. C., THOMPSON S. A., « Rhetorical Structure Theory : toward a functionnal theory of text organization », *Research Report RR-87-190, USC/Information Sciences Institute*, Marina del Rey, CA, 1988, p. 243-281.

[MAN 04] MANI I., *Narrative Summarization*, vol. 45/1, 2004.

[MAR 98] MARCU D., « Improving summarization through rhetorical parsing tuning », Montreal, Canada, 1998.

[MCK 01] MCKEOWN K. R., BARZILAY R., EVANS D., HATZIVASSILOGLOU V., SCHIFFMAN B., TEUFEL S., « Columbia Multi-Document Summarization : Approach and Evaluation », *in Proceedings of the Workshop on Text Summarization, ACM SIGIR Conference*, DARPA/NIST, Document Understanding Conference, 2001.

[MIN 04] MINEL J.-L., *Le résumé automatique de textes : solutions et perspectives*, vol. 45/1, 2004.

[OKA 01] OKA M., UEDA Y., « Phrase-representation Summarization Method and Its Evaluation », *in Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, mars 2001.

[ONO 94] ONO K., SUMITA K., MIIKE S., « Abstract Generation based on Rhetorical Structure Extraction », *in Proceedings of the 15 th International Conference on Computational Linguistics – COLING'94*, vol. 1, Kyoto, Japan, 1994, p. 344-348.

[RAD 98] RADEV D. R., MCKEOWN K., « Generating Natural Language Summaries from Multiple On-Line Sources », *Computational Linguistics*, vol. 24, n° 3, 1998, p. 469-500.

[RAD 04] RADEV D. R., JING H., STYŚ M., TAM D., « Centroid-based summarization of multiple documents », *DUC 2003*, décembre 2004, p. 919-938.

[SAL 73] SALTON G., YANG C., « On the specification of term values in automatic indexing », Journal of documentation 29, avril 1973.

[SID 02] SIDDHARTHAN A., « Resolving Relative Clause Attachment Ambiguities using Machine Learning Techniques and WordNet Hierarchies », *5th National Colloquium for Computational Linguistics in the UK (CLUK 2002)*, 2002, p. 45-49.

[TUR 03] TURNEY P., « Coherent Keyphrase Extraction via Web Mining », *in Proceedings Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, 2003, p. 434-439.

[WAN 03] WAN S., DALE R., DRAS M., PARIS C., « Straight to the Point : Discovering Themes for Summary Generation », *in Proceedings of the Australian Workshop on Natural Language Processing*, Melbourne, Australia, 2003.

compcir /
$FS = COMPCIR, SEMOBJ\ !=\ TEMPS \quad \wedge$
$PLACE\ !=\ ANT(TPH, INT) \quad \wedge$
$(\quad PLACE\ !=\ DROITE(TYP, VETAT/K, PHRASE)$
$\vee$
$PLACE = DROITE(FS, COMPCIR/K, PHRASE) \quad)$
gadj /
$FS = ATTR \quad \wedge$
$SOUSA = ADNOM \quad \wedge$
$SOUSATTR\ !=\ ATTRSUJ \quad \wedge$
$PLACE\ !=\ DROITE(SOUSD, ARTD/K, GN) \quad \wedge$
$PLACE\ !=\ DROITE(SOUSD, ARTD/K, GNPREP)$
phger /
$KPH = PHGER$
phrel /
$KPH = PHREL \quad \wedge$
$TYPREL\ !=\ OBJ$

(a) Les règles de sélection appliquées au texte de la figure 5.

Variable	Nom complet	Variable	Nom complet
FS	fonction syntaxique	SEMOBJ	sémantique de l'objet
TPH	type de la phrase	TYP	interprétation des construct. synt.
K	catégorie des groupes	SOUSA	catégorie de l'adjoint
SOUSATTR	type d'attribut	SOUSD	catégorie des déterminants
KPH	type de proposition	TYPREL	type du pronom relatif

(b) Définition des variables utilisées dans notre jeu de règles.

Valeur	Nom complet	Valeur	Nom complet
COMPCIR	complément circonstanciel	TEMPS	temps
INT	phrase interrogative	VETAT	verbe d'état
PHRASE	groupe phrase	ATTR	attribut
ADNOM	adjectif	ATTRSUJ	attribut du sujet
ARTD	article défini	GN	groupe nominal
GNPREP	groupe nominal prépositionnel	PHGER	proposition au gérondif
PHREL	proposition relative	OBJ	objet

(c) Définition des valeurs utilisées dans notre jeu de règles.

Figure 4.

MAUI PART À LA RECHERCHE DE SES PARENTS.
À 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 coeur 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.

Légende : compcir (complément circonstanciel), phger (proposition au gérondif), phrel (proposition relative), gadj (groupe adjectival).

Figure 5. *Coloration d'un texte, d'après notre méthode de compression de phrases*