Compression de phrases par élagage de leur arbre morpho-syntaxique

Une première application sur les phrases narratives

Mehdi Yousfi-Monod — Violaine Prince

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

RÉSUMÉ. *Nous proposons une technique de contraction de phrases qui se fonde sur l'étude de la fonction syntaxique et de la position dans l'arbre syntaxique des constituants des phrases. Nous définissons et analysons la perte de contenu et de cohérence discursive que la suppression de constituants engendre. Notre méthode de contraction s'oriente 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. Nous commentons les résultats obtenus sur un texte narratif court, totalement analysé, ce qui nous mène à poser le problème de l'évaluation quantitative des résultats d'une telle approche, par opposition à des résultats qualitatifs. La technique dépendant très fortement de la qualité de l'analyse, la question de la compression par élagage apparaît comme intimement subordonnée à l'évaluation de l'analyse syntaxique.*

ABSTRACT. *We propose a sentence compression technique which uses constituents syntactic function and position in the sentence syntactic tree. We analyze contents and discourse consistency losses caused by deleting such constituents. We explain why our method works best with narrative texts. With a rule-based system using SYGFRAN's morpho-syntactic analysis for French, we select removable constituents. We discuss the results obtained on a short narrative text, which has been completely analyzed. This rises the problem of quantitative versus qualitative evaluation. Since our technique is highly dependant on the quality of syntactic analysis, summarizing through pruning seems intimately intermeshed with parsing evaluation.*

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

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

1. Introduction

Le résumé automatique consiste à fournir à un ordinateur un document numérique afin qu'il en produise un nouveau document, plus petit, conservant les informations les plus importantes. Cette tâche pouvant s'effectuer à différents niveaux de qualité et d'objectifs, les types d'approches du résumé automatique sont 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 c'est 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. Aborder le résumé automatique sous toutes ses facettes, simultanément, est donc une tâche difficile. 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. Ses principaux apports sont fournis en section 2, qui traite de l'état de l'art. 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. Moins prolifique que la première, elle possède cependant un certain nombre de travaux fondateurs, dont les plus importants sont présentés et commentés dans l'état de l'art.

Enfin, une nouvelle démarche, issue du fait que les documents numériques sont de plus en plus structurés (et donc obéissent à des langages balisés), se fonde sur la récupération des balises comme indices de macrostructuration des textes permettant de les décomposer et donc de les contracter. Dans le cas de notre étude, nous nous intéressons uniquement au résumé de textes bruts, et donc non balisés. Par conséquent, nous ne nous attarderons pas sur le résumé fondé sur la récupération des balises.

---

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, soit le groupe nominal « un médecin de famille ». Il est composé de deux constituants : un groupe nominal « un médecin » et un groupe nominal prépositionnel « de famille ». Ce dernier a un rôle de modificateur du premier.

1.1. *Définition des points-clés de notre démarche*

Produire un résumé pour un être humain nécessite un effort cognitif élevé. Il est difficile de faire effectuer cette tâche aux ordinateurs car, à l'heure actuelle, leurs capacités cognitives sont encore très inférieures à celles des humains. Une variante à cette tâche consiste à supprimer les phrases les moins pertinentes d'un document : le résumé par extraction de phrases-clés (type de résumé nommé *extracts*). Cette technique est plus abordable par un ordinateur, elle se retrouve d'ailleurs majoritairement utilisée dans les travaux actuels. Un de ses inconvénients, cependant, est qu'aucun traitement n'est effectué au niveau intra-phrase, ainsi une longue phrase est soit conservée dans son intégralité, soit totalement supprimée : il n'y a pas d'intermédiaire. La compression de phrase peut alors combler ce manque en pénétrant dans les phrases afin d'en supprimer les constituants les moins pertinents. L'idée centrale de notre recherche est de traquer les limites de la contraction de textes (type de résumé nommé *abstract*) par compression de phrases sans perte majeure d'information et sans perte de cohérence grammaticale.

1.1.1. *Perte de cohérence grammaticale*

Un constituant perd la **cohérence grammaticale** si un sous-constituant *gouverneur* de ce dernier est supprimé. Par exemple, dans le groupe nominal prépositionnel « une maison de campagne », « maison » est gouverneur, s'il est supprimé, le groupe est incohérent. En revanche, « campagne » est modifieur, et apparaît comme un sous-constituant *incident*, dont la suppression entraîne certes une perte d'information, mais qui ne nuit pas à la compréhension. Dès lors, une phrase, qui est une composition de constituants, est considérée comme ayant subi une incohérence grammaticale, si elle perd un ou plusieurs constituants gouverneurs. Dans la théorie de Chomsky (Chomsky, 1982) sont **gouverneurs** les constituants dont le rôle syntaxique est sujet ou prédicat, en d'autres termes le groupe sujet, et le verbe principal d'action. Les autres constituants sont des modifieurs de verbe (compléments d'objet et certains compléments circonstanciels) ou modifieurs de phrase (essentiellement des compléments circonstanciels). Leur suppression pourra être envisagée de façon graduelle, et c'est précisément le sujet de l'étude et de son expérimentation.

1.1.2. *Perte majeure d'information*

Par **perte majeure d'information**, ou perte de contenu important, nous entendons la chose suivante : un constituant qui reste grammaticalement cohérent, perd une information majeure si la suppression du ou des sous-constituants incidents :

– annule ou affaiblit la complétude syntaxique. Par exemple, le passage d'un mode transitif à un mode intransitif pour un prédicat produit une perte majeure d'information. « Je mange des prunes » peut être réduit à « Je mange » sans incohérence, mais la complétude du prédicat est affaiblie. « Je donne des prunes », réduit à « Je donne » peut devenir incompréhensible : la complétude est annulée ;

– nuit à la cohésion sémantique du texte. Chaque phrase peut rester grammaticalement cohérente, voire syntaxiquement complète, mais l'ensemble des phrases peut apparaître décousu car un élément pivot a été supprimé. Nous chercherons justement par l'expérience à ajuster la notion de contenu important par rapport aux théories linguistiques existantes, et à l'objectif du résumé.

1.2. *Organisation de l'article et principal positionnement*

Une fois que les avantages et les inconvénients de cette méthode auront été discutés, nous passerons à la compression de texte fondée sur des repères de plus grande granularité (voir section 3.2) que celui de sous-constituant d'un constituant, tout en restant dans les limites de la phrase. En effet, il importe d'isoler les différentes variables concourant à la contraction de texte afin d'en déterminer l'impact. La véritable originalité de notre démarche 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. Nous verrons par la suite que 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 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 ou *extracts*, puis nous comparons ceux qui travaillent à un niveau de granularité plus fin (section 2) ; nous présentons ensuite notre méthode de contraction de textes (*abstract*), fondée sur la compression de phrases à partir de leur analyse morpho-syntaxique (section 3). Nous illustrons ensuite notre proposition à l'aide des sorties d'une application prototype appliquée à un texte du genre conte (section 4).

Il est très clair que l'application prototype est un premier stade d'expérimentation et ne peut en aucun cas jouer un rôle d'évaluation dans le sens où ce dernier est actuellement compris dans la littérature. Cela pour plusieurs raisons :

– les *abstracts* fondés sur une approche syntaxique sont peu abordés dans le domaine en raison du faible nombre d'analyseurs en dépendances complets et robustes, les méthodes d'évaluation ne sont donc pas encore adaptées à ce type d'approche. Justement, une évaluation des analyseurs du français a eu lieu dans la campagne EVALDA/EASY, mais les difficultés rencontrées par les évaluateurs ne sont pas négligeables. Comment comparer entre eux des analyseurs n'obéissant pas aux mêmes impératifs grammaticaux, ou comment corriger les évaluations humaines servant de référence ? A ce titre, une discussion intéressante a lieu autour de ce thème, que nous reprendrons dans notre section 5, car elle place exactement à sa juste mesure, l'importance que l'on peut donner à l'évaluation quantitative. Une évaluation des *abstracts* produits faite par comparaison ne peut que difficilement se faire, pour les raisons suivantes :

- la faible fréquence des travaux de ce type, par rapport aux travaux sur les extracts ;

- la diversité des méthodes de production d'*abstracts* qui ont peu de points sur lesquels asseoir les comparaisons ;

- la difficulté de faire de l'évaluation quantitative dès qu'on rentre dans le domaine de la syntaxe. Certes, des travaux comparatifs existent (Lin, 2003) mais ils sont davantage orientés vers des mesures relatives sur un corpus donné, que sur une réflexion de fond sur la méthode. Or, nos premières observations montrent que la méthode doit être paramétrée par la nature du corpus. Cela signifie que l'on n'adoptera pas les mêmes règles de contraction pour deux corpus de caractéristiques différentes, ces dernières étant à déterminer ;

– la discussion sur la validité des méthodes syntaxiques (Lin, 2003) est parfaitement biaisée. Nous en discutons en section 2.2.5 ;

– l'évaluation automatisable n'est pas facile pour des *abstracts*. Si leurs produits duaux, les *extracts*, peuvent être évalués par des tâches de type question-réponse (à la TREC), il ne reste plus, puisque nous réfutons la mesure sur corpus unique car elle n'est pas probante, que l'évaluation par jugement humain. De ce fait, nous montrons dans la section 5 que, pour offrir à ce dernier des éléments probants utilisables dans la réalisation d'un résumeur complet, il faut, à partir de cette première expérience relatée dans cet article, définir un protocole rigoureux de paramétrage de la procédure. Nous en donnons les prémices dans la section en question.

2. L'état de l'art en matière de résumé automatique

Une grande variété de techniques sont utilisées allant de la production d'un résumé par extraction à celle d'un 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. Nous présentons un panorama général ici afin que la place des *abstracts* puisse être mieux appréhendée par les lecteurs non spécialistes du domaine.

Les approches abordant le résumé par reformulation (Radev *et al.*, 1998; McKeown *et al.*, 2001; Daumé III *et al.*, 2002; Oka *et al.*, 2001) sont souvent basées sur l'utilisation de structures de données intermédiaires (patrons dans (Radev *et al.*, 1998), graphes de relations dans (Oka *et al.*, 2001)...) remplies avec des informations extraites du texte source puis fournies en entrée à des outils de génération de langue (comme FUF/SURGE de (Elhadad *et al.*, 1996)) qui produisent le texte résumé final. Ces approches ayant un mode de production s'éloignant grandement du nôtre, nous ne les détaillons pas davantage dans cet article.

Les méthodes par extraction sont fondées sur l'hypothèse *qu'il existe, dans tout texte, des unités textuelles saillantes* (Minel, 2004). 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 cherche à repérer ces unités saillantes et propose un texte de taille plus petite que le document initial qui garde majoritairement ces unités. Nous faisons également l'hypothèse de l'existence de ces unités, ainsi que de leur intérêt pour le résumé. Ce seront les constituants **gouverneurs**, décrits en section 3.2, qui correspondront à ces unités saillantes.

2.1. *Extraction de phrases-clés*

La plus grande partie des approches du résumé de texte procède par extraction de phrases-clés (production d'*extracts*), le but étant de choisir les meilleures candidates et de les placer bout à bout pour produire le résumé final. Plusieurs de ces approches s'appuient sur des techniques statistiques, dans lesquelles des informations basées sur la fréquence des termes, comme le produit *tf*idf* de (Salton *et al.*, 1973), sont fréquemment utilisées pour évaluer l'importance de chaque phrase dans un document. Par exemple, les travaux relatés dans (Luhn, 1958; Barzilay *et al.*, 1997; Goldstein *et al.*, 2000; Boguraev *et al.*, 2000; Lin *et al.*, 2002; Radev *et al.*, 2004; Erkan *et al.*, 2004) l'utilisent.

De nombreuses autres techniques sont aussi employées :

– les méthodes probabilistes de catégorisation, souvent assorties d'un moteur d'apprentissage (comme les modèles de Markov), et se fondant sur un corpus de documents associés à leur résumé (Julian *et al.*, 1995; Turney, 2003) ;

– les méthodes utilisant des espaces vectoriels : (Ando *et al.*, 2000) utilise une technique proche de la décomposition en valeurs singulières (*Singular Value Decomposition, SVD*) de l'indexation sémantique latente (*Latent Semantic Indexing, LSI*) (Deerwester *et al.*, 1990), ou encore (Hirao *et al.*, 2002) se fonde sur les *Support Vector Machines* pour séparer les phrases-clés des autres ;

– les chaînes de coréférence : (Baldwin *et al.*, 1998; Azzam *et al.*, 1999) ;

– les chaînes lexicales : (Chaves, 2001; Fuentes *et al.*, 2002; Alemany *et al.*, 2003), ces approches ont tendance à prendre comme unité textuelle le paragraphe plutôt que la phrase ;

– l'utilisation de la structure rhétorique (Mann *et al.*, 1988) comme (Ono *et al.*, 1994) 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'avantage de ces approches est qu'elles sont relativement bien adaptées au résumé thématique (c'est-à-dire, orienté par le thème, ce qui lui donne sa valeur d'*extract*) et non pas un résumé image fidèle du texte, qui correspond à la définition de l'*abstract*. Leur inconvénient est que la structuration même des phrases ainsi sélectionnées n'est pas toujours compatible avec ce que l'on attend d'un résumé. Nous verrons par la suite

que dans le cas de certains types de texte (romans, contes...), les phrases peuvent être longues et posséder des informations non indispensables à la compréhension globale. 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 trop grandes d'une part, et d'autre part, de ne pas chercher à supprimer systématiquement 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 rhétorique (qui pourrait s'appuyer sur des textes balisés avec des langages de présentation comme XML) ;

– et enfin 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 ce texte.

(Oka *et al.*, 2001) 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 dans le graphe afin d'ajouter des 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 *et al.*, 2003) 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 peut parfois être mise en cause, et dans lesquelles la perte majeure d'information est possible, alors que du superflu peut être incidemment conservé.

2.2.2. *Le « copier/coller »*

(Jing *et al.*, 2000) utilise des phrases-clés sélectionnées par des résumeurs classiques, les comprime, puis les combine en de nouvelles phrases (la méthode est détaillée à la section 2.2.4). A partir d'un corpus de documents associés à leur résumé réalisé par des experts, les auteurs ont déterminé six opérations qui permettent (utilisées seules ou combinées) de transformer une phrase en une phrase compressée, avec pour objectif de se rapprocher le plus possible des phrases résumées par les humains. Les différents types d'opérations sur les phrases incluent la réduction, la combinaison, la transformation syntaxique et le paraphrasage lexical.

(Ishikawa *et al.*, 2002) 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 deux 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*

(Marcu, 1998) 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 deux nœuds 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, ce qui tend à renforcer l'idée qu'un

corpus ayant un genre donné, comme unité d'évaluation n'est pas discriminant. L'auteur n'est pas parvenu à trouver une solution fonctionnant pour tout genre de texte. Il aurait pu se préoccuper de rechercher des intervalles de valeurs pour calibrer son système, mais il n'a malheureusement pas poussé la discussion jusque là.

Une fois la SR déterminée, un ordre partiel entre les différents satellites est établi, les plus proches de la racine se voient attribuer une importance plus grande. Les satellites sont ensuite supprimés, des moins importants aux plus importants selon la taille du résumé désirée. La cohérence est assez bien conservée dans les cas où l'analyse de la SR est correcte. Si les idées fondamentales de cette technique sont très proches de celles sur lesquelles nous nous appuyons, sa principale difficulté reste de détecter correctement la SR, ce qui supposerait l'existence d'un analyseur « rhétorique », chose encore plus difficile qu'un analyseur en dépendances.

2.2.4. *La compression de phrases*

(Grefenstette, 1998) utilise la nature des syntagmes et propositions pour estimer leur importance, puis supprime les moins importants pour produire les phrases compressées. La cohérence obtenue est évidemment faible mais suffisante pour l'application souhaitée qui est la réduction de textes télégraphiques destinés à être lus pour les malvoyants.

(Knight *et al.*, 2002), qui est justement l'approche critiquée par (Lin, 2003) et qui à première vue pourrait se rapprocher de la nôtre, aborde le problème sous deux angles : un modèle probabiliste et un modèle basé sur la décision. Les deux modèles utilisent le parseur Collins (Collins, 1997) qui génère des arbres syntaxiques des phrases d'un texte et donne des informations sur la nature des syntagmes de ces phrases.

Le premier emploie un modèle de canal bruité (*noisy-channel model*) qui consiste à faire l'hypothèse : *la phrase à comprimer fut autrefois courte et l'auteur y a ajouté des informations supplémentaires (le bruit)*. Le but est alors de retrouver ces informations pour les supprimer. Le modèle probabiliste est bayesien, et les auteurs l'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 est attribuée à une phrase comprimée lorsque cette dernière est incorrecte grammaticalement ou a perdu certaines informations comme la négation. Pour réaliser leur évaluation, les auteurs ont créé un corpus de test en extrayant 32 paires de phrases (phrase originale, phrase résumée) de leur corpus. Les autres paires de phrases (au nombre de 1035) constituaient le corpus d'entraînement. Leur métrique est fondée sur un score de bi-grammes de caractères. Les résultats sont assez concluants, cependant la justesse grammaticale n'est pas suffisamment bien conservée dans la plupart des cas et une légère perte d'« information importante » est à noter.

Le second modèle utilise des règles de transformations appliquées aux arbres syntaxiques des phrases du texte dans le but de réduire ces arbres puis de recomposer des phrases plus courtes. Les règles sont composées d'un ensemble d'opérations élémentaires de manipulation des arbres. Le moteur d'apprentissage chargé de créer

ces règles utilise le programme C4.5 (Quinlan, 1993). Cette approche tient compte de la nature des constituants et leur position dans l'arbre syntaxique, mais leur fonction syntaxique n'est pas prise en compte. Le taux de compression n'est pas paramétrable et avoisine les 55 % dans l'expérimentation. La cohérence grammaticale et la conservation de contenu important (l'absence de perte majeure d'information) sont légèrement inférieures au premier modèle des auteurs.

(Jing, 2000) compresse les phrases en utilisant :

– un corpus contenant des phrases et leur forme compressée correspondante, écrite par des humains ;

– un lexique incluant des sous-catégorisations de verbes, utilisé pour déterminer les arguments indispensables des verbes des propositions ;

– la base de données lexicale Wordnet (Miller *et al.*, 1990) contenant des relations lexicales (synonymie, antonymie, méronymie...) entre les mots ;

– le parseur *English Slot Grammar* (ESG) (McCord, 1990) qui annote les constituants des phrases avec leur nature et leur rôle thématique.

Voici les différentes étapes de l'algorithme :

1) le parseur ESG produit l'arbre syntaxique des phrases ;

2) les constituants indispensables à la cohérence grammaticale sont déterminés à l'aide des informations sur leur rôle thématique et sur les sous-catégorisations des verbes les précédant dans la phrase ;

3) le système décide quels sont les constituants les plus proches du thème du document à l'aide de Wordnet : les relations lexicales entre les mots sont extraites, plus la connexité d'un mot avec les autres est élevée, plus il est considéré comme proche du thème du texte ;

4) une probabilité d'être supprimé est attribuée à chaque constituant des phrases, en utilisant le modèle de Bayes, à partir du corpus de phrases et leur version compressée, en fonction des verbes utilisés et du rôle thématique des constituants ;

5) la réduction des phrases est ensuite effectuée en fonction des annotations précédentes, après une pondération des différents facteurs d'importance et de cohérence, et un seuil fixé en fonction de la qualité du résumé désirée.

L'évaluation de leur système, faite par les auteurs, a montré que les choix de suppression des constituants concordaient à peu près à 81 % avec ceux faits par des humains.

La taille des textes est réduite d'environ 33 % par le système, contre 42 % par des humains. Cette approche a l'avantage de pondérer l'importance des constituants par des données contextuelles (étape 3), informations dont nous ne tenons actuellement pas compte dans notre approche. La cohérence grammaticale est assurée par l'étape 2. La perte majeure d'information est fortement dépendante de l'étape 4 qui se base sur un modèle d'apprentissage probabiliste utilisant le rôle thématique des constituants pour déterminer leur importance. Notre approche diffère en ce point d'une part

en ce qu'elle se base sur un système de règles créées manuellement (modèle non probabiliste), d'autre part en ce qu'elle utilise la fonction syntaxique et donc le rôle syntaxique des constituants des phrases.

2.2.5. *Critique (contestable) du résumé par compression de phrases*

(Lin, 2003) a évalué la qualité d'un résumé produit par extraction de phrases-clés puis compression des phrases extraites. La méthode d'extraction des phrases est celle de (Lin *et al.*, 2002) et la méthode de compression est celle de (Knight *et al.*, 2000), basée sur le modèle de canal bruyant (similaire à (Knight *et al.*, 2002)). L'auteur 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 l'auteur n'utilise que la méthode de (Knight *et al.*, 2000) pour comprimer les phrases, nous contestons la généralisation de sa conclusion à l'ensemble des méthodes de compression. La seule conclusion possible est que la méthode de compression utilisée, qui, en pratique, mélange à la fois statistiques, apprentissage, technique du canal bruité et structure syntaxique, ne satisfait pas aux contraintes de cohérence grammaticale et de conservation du contenu. Notre approche diffère grandement de celle de (Knight *et al.*, 2000) sur au moins un point : nos règles de compression sont produites manuellement, en relation avec des modèles linguistiques, puis mises en œuvre, et non inférées automatiquement de façon calculatoire.

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

Le point de départ de notre approche fut l'intuition que, pour la compréhension d'un texte, **la fonction syntaxique et la position dans l'arbre syntaxique des constituants des phrases sont deux facteurs conséquents dans l'évaluation de l'importance de ces constituants**. Cette intuition prend ses racines dans l'analyse grammaticale logique classique et dont on trouve des manuels connus (Mauffrey *et al.*, 1995; Grevisse, 1993-1997; Wagner *et al.*, 1962). En effet, ne sont pas toujours indispensables pour comprendre le sens principal de la phrase, certains épithètes, certains compléments circonstanciels, etc.

Par exemple, dans la phrase « *Un chat gros et laid mange une souris* », le groupe adjectival coordonné « gros et laid » peut être supprimé sans nuire réellement à la compréhension (il n'est pas gouverneur, cela reste donc grammaticalement cohérent), et à l'intérêt (la phrase ainsi contractée ne nuit pas *a priori* à la cohésion sémantique du texte réduit).

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 section 3.4 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 (Chauché, 1984). SYGFRAN utilise un ensemble de règles de transformations d'éléments structurés, mettant en œuvre les règles de la grammaire française[2], 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 suivants :

– la rapidité : la complexité théorique d'analyse est en $O(k * n * log_2(n))$ où k est le nombre de règles (12000 en décembre 2005) et n la taille du texte en nombre de mots. 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 (en fait il est de l'ordre de 16). Cela dit, même ainsi, plus le texte est important, plus k est petit devant n. Aujourd'hui SYGFRAN analyse un corpus de 220000 phrases, d'en moyenne 25 mots, en environ 24 heures, sur un ordinateur grand public disposant d'un processeur cadencé à 2,4 Ghz et d'une capacité de mémoire vive de 1 Go ;

– la robustesse : SYGFRAN parvient, en décembre 2005, à obtenir une structure correcte pour au moins 35 % 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.

Les noms des nœuds 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 PREPositionnel et $GCARD$ pour Groupe CARDinal. Les noms des feuilles (ellipses) sont les formes canoniques des lexies (masculin, singulier, infinitif). Le numéro de chaque nœud est un pointeur sur les informations des variables SYGFRAN associées au nœud. Par exemple, le nœud 3 possède entre autres les variables et valeurs suivantes :

2. La grammaire choisie est celle définie par (Grevisse, 1993-1997), en relation avec les travaux du linguiste Jürgen Weissenborn. Le choix a été fait par l'auteur de l'analyseur.

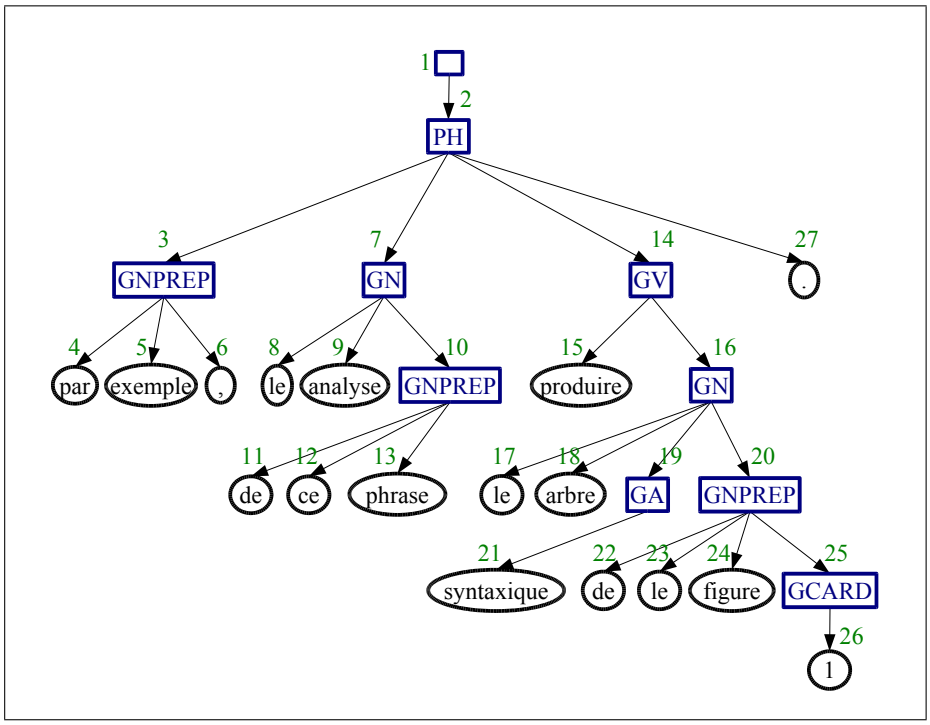

Figure 1. *Exemple d'analyse de SYGFRAN*

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	le rôle syntaxique est « complément circonstanciel »

Le nœud 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 nœud 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.

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 »

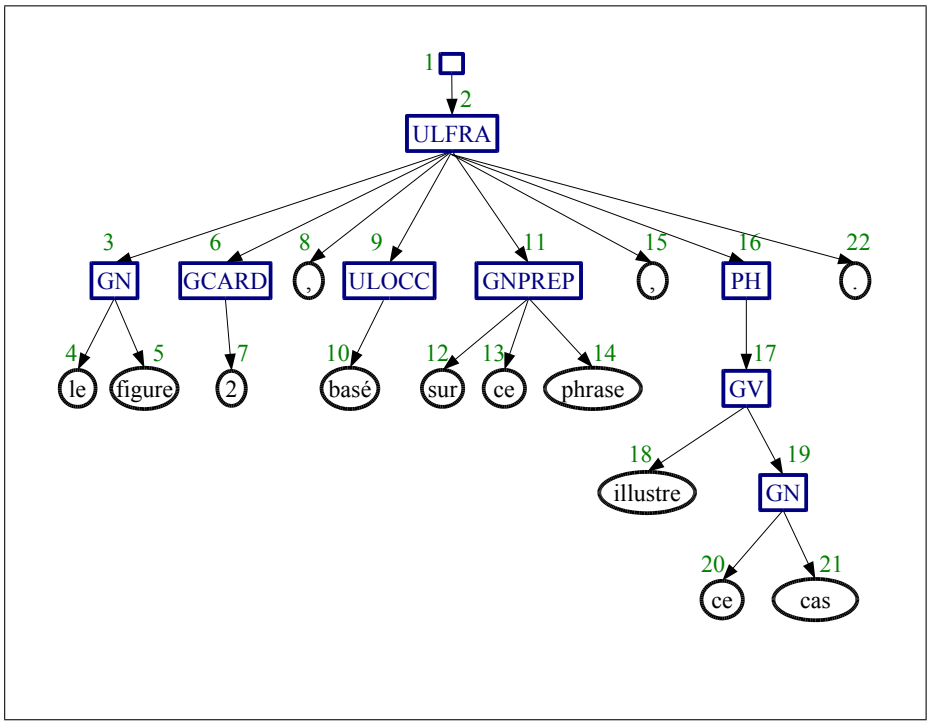

Figure 2. *Exemple d'analyse ayant partiellement échoué*

(*Unknown Locution*) comme nom du nœud 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 nœud 11, contient l'information que « sur cette phrase » est un groupe prépositionnel.

3.2. *Fonction, rôle et position*

Le test de suppression des constituants est abordé par de nombreux ouvrages sur la grammaire française, dont ceux cités au début de cette section[3], 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 présentées par les linguistes ((Tomassone, 2001) pour les compléments circonstanciels par exemple), mais pas de

3. Il est aussi appelé *effacement* ou *soustraction*. En voici une définition : manipulation syntaxique qui consiste à supprimer un mot ou un groupe de mots. On l'appelle aussi la soustraction.

règle fondamentale. Nous avons donc considéré ces recommandations comme des hypothèses de travail et nous avons cherché à les étayer empiriquement. Comme abordé en section 1, sont **gouverneurs** au sens de Chomsky (Chomsky, 1982) des constituants considérés comme indispensables à la cohérence grammaticale et sémantique de la phrase. Le sujet d'une proposition (verbale) et son prédicat d'action sont gouverneurs. Si un gouverneur ne peut être supprimé, en revanche, il faut chercher à le réduire si cela est possible, en préservant toujours la cohérence grammaticale, et sans perte majeure d'information.

Considérons la phrase simple[4] suivante : « Jean mange une pomme verte ». Pour le maintien de la cohérence grammaticale, on ne peut supprimer aucun gouverneur (sujet et verbe). On conserve donc « Jean », le sujet. Comme il est atomique, on ne peut pas le réduire. Le groupe verbal comprend « mange une pomme verte ». En tant que groupe, il est gouverneur, et donc ne peut être supprimé. Voyons s'il peut être réduit. Dans ce groupe, le verbe « mange » est gouverneur et atomique. Il doit être conservé. 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 l'on affaiblit la complétude syntaxique (voir définition en section 1) : 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 complétude syntaxique (pas de modification du mode du prédicat).

Ainsi la détermination du constituant *secondaire* ou *incident* 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 nominal**.

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 circonstancielles attachées à la proposition principale : « Jean mangera une pomme verte quand la saison de la cueillette arrivera. »

– les propositions relatives tenant lieu de modifieur d'un groupe nominal complément d'objet : « Jean mange une pomme verte qu'il a cueillie sur le premier pommier de son verger. »

– 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. »

4. Réduite à une proposition principale.

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[5].

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 majeure d'information. Tout le problème consiste à savoir si :

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

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

– on peut élaguer des constituants nominaux ;

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

3.3. *Type d'effacement et type de texte : la dissimilarité des corpus*

Pour cela, à partir de textes de genres variés, nous avons réalisé des tests de suppression de certains constituants en fonction de leur fonction ou de leur rôle syntaxique (donc plutôt la granularité « moyenne »), en estimant, parfaitement qualitativement, les pertes de cohérence discursive et de contenu important dans les phrases comprimées. Ces premiers tests étant essentiellement pour se faire une idée, le protocole de constitution des corpus n'était pas particulièrement important. En substance nous avons considéré :

---

5. (Tomassone, 2001) dit spécifiquement : *Dans la tradition scolaire, les compléments circonstanciels sont plus ou moins explicitement posés comme non-essentiels*. Elle préfère le terme de circonstant à celui de complément circonstanciel traditionnel, car certains sont justement essentiels, comme nous le verrons dans le paragraphe 3.3. Elle définit le circonstant comme étant facultatif, ou effaçable.

– le corpus de dépêches traitées par SYGFRAN pour la catégorisation dans (Chauché *et al.*, 2003) (220 000 phrases, de style hétérogène et journalistique) ;

– quelques articles scientifiques en biologie ;

– quelques textes narratifs (romans, contes...).

Nous avons fait des « sondages » en prenant au hasard des textes ou des sous-textes, et nous avons essayé la méthodes d'effacement, à la main, avant même d'avoir à l'évaluer automatiquement, pour savoir si la méthode était capable de résister à une première « plongée » dans les textes. Les constatations que nous avons faites, à partir des données considérées sont les suivantes.

Dans les textes du genre article scientifique ou énoncé technique, chaque constituant se révèle avoir beaucoup plus d'importance que dans un texte narratif. Prenons par exemple le terme « hormone de synthèse » (article de biologie), 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 km/h soufflera sur le Golfe du Lion* » de son complément circonstanciel de lieu (« le Golfe du Lion »), dans les dépêches météo. 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 cœur de cette dernière. 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 intra-phrastique, 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 des formes plus macroscopiques de compression pour les textes scientifiques, techniques ou journalistique.

(Mani, 2004) aborde la problématique du résumé de textes narratifs, en s'appuyant principalement sur des indices temporels. Il étudie les événements sur trois plans : la scène, l'histoire et l'intrigue, dans le but d'extraire les événements-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, puisque ce sont celles que nous espérons obtenir automatiquement à partir de SYGFRAN.

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 en 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.

L'étude de la fonction syntaxique nous a amené à noter trois catégories principales de constituants susceptibles d'être supprimés : les compléments circonstanciels (section 3.3.1), les épithètes (section 3.3.2) et les appositions (section 3.3.3). Comme on peut le voir, ils sont de granularité moyenne. Toutefois il existe des cas de granularité plus importantes, notamment pour les propositions relatives en appositions, qui peuvent jouer le rôle d'épithètes. Leur suppression augmente le taux de compression obtenu.

3.3.1. *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. Une 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. Ces trois compléments ne peuvent, dans un certain nombre de cas, être supprimés. Ceux de *manière* (questions « *Comment ?* ») et de *cause* (questions « *Comment est-ce arrivé ?* ») sont peu importants dans une majorité des cas. (Tomassone, 2001) donne une règle intéressante pour l'importance de leur effacement : *sont considérés comme effaçables, les circonstants qui n'affectent ni la négation, ni l'interrogation.* En effet, un CC de lieu situé dans une phrase interrogative est important car la question porte généralement sur lui. Exemple : « *Jean n'aurait-il pas dormi dans la confiserie ?* » ce qui montre que la sensibilité à la négation ou à l'interrogation (dans l'exemple, les deux sont mêlés) est un bon test pour la conservation du CC.

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. Il est clair que la théorie linguistique considérée porte sur les verbes d'action. Le problème des verbes d'état est très différent. En effet un groupe ayant une sémantique de lieu placé après un verbe de ce type, fait partie du prédicat, et ne peut donc être supprimé[6]. Par exemple, on ne peut supprimer « *dans la voiture* » de la phrase « *Jean est dans la voiture* ». Cependant, si plusieurs groupes locatifs se suivent après le verbe d'état, 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* ». En effet, les autres

6. Les auteurs remercient tout particulièrement Augusta Mela, maître de conférences en sciences du langage, pour l'avoir fait remarquer.

groupes sont des *circonstants de phrase* (phrase dont le prédicat gouverneur est « *être dans la voiture* »), et de ce fait, d'après notre nomenclature, et en accord avec les propositions de (Tomassone, 2001), ils sont facultatifs, et donc peuvent *a priori* être supprimés.

3.3.2. *Les épithètes*

Certains 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. Par ailleurs, 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. Cela est dû au fait que l'article défini est utilisé pour parler d'une entité particulière et que les épithètes du nom permettent 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 cause une perte de contenu considérable (cohésion sémantique menacée). En revanche, dans la phrase « *Il vit un enfant blond dans la rue* », l'adjectif est moins important et peut donc être supprimé.

3.3.3. *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 3.3.2) ;

– 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 n'a qu'une crainte, manger des légumes* »).

Dans les trois premiers cas, les constituants se suppriment sans difficulté. Les propositions participales sont aussi de bons candidats 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 une information importante qui vient compléter le sujet.

3.4. *Architecture*

L'architecture de notre système est présentée 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. Mise en œuvre de la méthode

Nous avons réalisé un programme prototype écrit en Java[7] afin de pouvoir étudier le domaine de compétence de cette approche aux caractéristiques fortement linguistiques. Nous avons défini un système utilisant des règles simples, basées sur les résultats de notre « étude théorique » (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 qu'il soit sélectionné puis supprimé. 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 nœud 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).

Pour avoir une estimation du passage d'un texte brut à un premier niveau de compression, 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. 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 4a), utilisant les variables décrites dans la figure 4b et les valeurs décrites dans la figure 4c. La première règle peut être traduite par : je nomme et sélectionne $compcir$ (CC) tout constituant qui :

7. `http://java.sun.com`

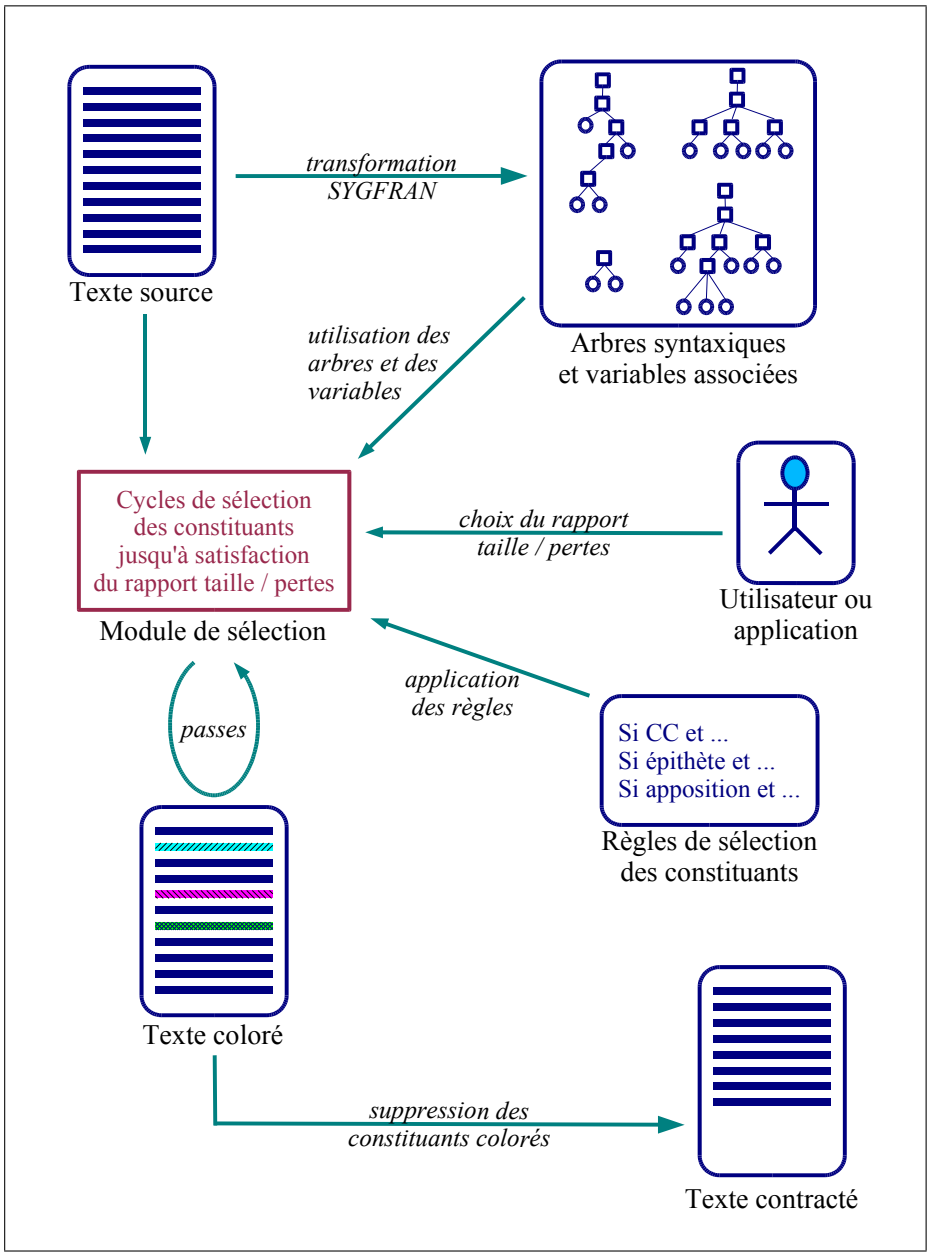

Figure 3. *Fonctionnement de notre système de compression de phrases*

– 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) ;

– vérifie soit :

- n'est pas situé à droite d'un constituant qui a « verbe d'état » comme interprétation des constructions syntaxiques, relativement à un nœud 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 nœud 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 : ce constituant est un complément circonstanciel mais pas de temps, il n'est pas dans une phrase interrogative ni n'est situé à droite d'un verbe d'état.

Nous avons utilisé comme texte de test un conte tahitien. La principale raison de ce choix est que SYGFRAN produit une syntaxe correcte pour l'intégralité des phrases de ce texte, supprimant par là le problème du traitement des analyses partielles. Le résultat de la coloration de la première moitié de ce texte est présenté en figure 5. Pour des raisons pragmatiques d'impression monochrome, les constituants sont soulignés ou en italique plutôt que d'être colorés (se référer à la légende en fin de figure).

4.1. *Discussion sur les résultats sur un texte de type « conte »*

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

4.1.1. *Aspects relatifs à notre prototype*

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. Les règles actuelles de sélection des constituants sont simples et peu nombreuses, elle peuvent donc être affinées, la principale contrainte étant que les données linguistiques dans ce domaine sont très limitées. L'affinage pourra porter sur la fonction des constituants et surtout

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	interprét. 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	PHREL	proposition relative
PHGER	proposition au gérondif	OBJ	objet

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

Figure 4. *Règles, valeurs et variables utilisées dans notre expérimentation*

selon le genre des textes. Nous comptons, à cet effet, effectuer des expérimentations sur plus de textes touchant à des genres plus variés.

MAUI PART A 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*.

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*

4.1.2. *Aspects relatifs à l'analyseur*

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. Evaluation et extension de la démarche

5.1. *Evaluation quantitative sur textes narratifs*

Si en pratique, la mise en œuvre du prototype sur un conte a permis de déceler un certain nombre d'avantages et de défauts, pour évaluer ce dernier sur ce type de textes, il faudrait réunir un corpus conséquent de documents narratifs. De par sa nature même, notre analyse est fortement dépendante de la qualité de l'analyse morpho-syntaxique en constituants et en dépendances. L'analyseur que nous employons, SYGFRAN, d'une part n'analyse correctement qu'environ 35 % de phrases prises au hasard dans des corpus hétérogènes, et d'autre part, parmi ces phrases, ne renseigne pas systématiquement avec justesse toutes ses variables (comme vu en section 4.1). Face à cette difficulté, plusieurs pistes ont été envisagées, puis rapidement abandonnées en raison de leur inefficacité.

5.1.1. *Piste 1 : analyse partielle de SYGFRAN*

La première et plus simple a été de chercher à évaluer le prototype avec les analyses partielles retournées par SYGFRAN. Malheureusement ces dernières introduisent des biais gênants pour l'évaluation : ainsi, nous avons constaté que des segments entiers, qui auraient pu être élagués, ont été conservés, simplement parce que la sous-arborescence concernée n'a pu être entièrement analysée. Pire encore, dans les parties non complètement analysées, nous avons vu apparaître de fausses étiquettes : les constituants n'étaient que partiellement cernés, et de fait, les dépendances, totalement escamotées. Nous avons fait un test sur un texte journalistique d'un millier de mots, en analyse « tout venant » : la très grande majorité des mauvaises compressions de phrases relevées dans nos tests était due à l'incomplétude des résultats de SYGFRAN.

5.1.2. *Piste 2 : autres analyseurs*

Une seconde solution aurait été d'utiliser un autre analyseur. Malheureusement, et en cela la campagne EASY d'évaluation des analyseurs syntaxiques du français l'a

montré :

– les analyseurs existants ne s'accordent entre eux que sur des constituants simples (groupe nominal, verbal...) et non pas sur des constituants plus complexes (propositions relatives, coordonnées...). L'ensemble de règles mis en place avec un analyseur ne vaut pas pour un autre ;

– la majorité des analyseurs autres que SYGFRAN ne produisent que très peu de dépendances[8]. En outre, ces dépendances sont superficielles ;

– SYGFRAN est le seul analyseur robuste (acceptant du tout venant) qui produise des arbres et non pas des listes d'analyse « à plat ». Par conséquent, toutes les règles se fondant sur les profondeurs des racines des sous-arborescences (pour exprimer la localité du groupe en question) ne peuvent s'appliquer ;

– enfin, la précision des autres analyseurs n'a pas été démontrée comme forcément meilleure que celle de SYGFRAN, sur les rares aspects évaluables en commun.

Autant dire que les « non résultats » d'EASY nous ont ancrés dans notre choix de SYGFRAN.

5.1.3. *Piste 3 : analyse manuelle*

La solution qui consiste à produire manuellement une analyse morpho-syntaxique à partir d'un corpus de textes n'est réalisable que pour un tout petit texte, et donc ne peut pas servir de caution quantitative pour l'évaluation. En effet, cette tâche est extrêmement difficile et fastidieuse, car dresser l'arbre syntaxique correct pour chaque phrase, tout en renseignant chaque variable utile à notre analyse, c'est-à-dire les fonctions syntaxiques révélant les relations entre constituants, demande un temps totalement inaccessible dans le cadre de notre travail. De plus, il doit être réalisé par un linguiste, qui utilise la même grammaire que SYGFRAN (celle de J. Weissenborn). Dans l'évaluation EASY, bien que les corpus proposés à l'analyse aient été conséquents, seule une petite partie des données a été étiquetée par des experts, et encore, uniquement en constituants. Entre les différents participants, les critiques sur les affectations des étiquettes, ainsi que la contestation des choix, a été constante : il n'existe pas une théorie grammaticale unique et unifiée du français (ni même de toute autre langue).

5.1.4. *Piste 4 : corpus existant*

Enfin une dernière solution aurait été d'utiliser des corpus existants déjà marqués par des informations syntaxiques. Par exemple le corpus EASY annoté à la main possède quelques informations syntaxiques. Cependant il est composé de structures « à plat » (aucune profondeur). Les dépendances sont très succintes (modifieur de GN ou de GV). De plus les organisateurs précisent, dans le « Guide d'annotation version

8. Francis Brunet-Manquat, dans sa thèse (Brunet-Manquat, 2004), utilise l'analyseur de Xerox. Il développe une plateforme DepAn pour une analyse en dépendances. Seules trois règles de dépendances ont été définies, par opposition aux dizaines introduites dans SYGFRAN.

1.6 » (Gendner *et al.*, 2004), qu'ils font *l'hypothèse qu'on n'a pas de constituants discontinus, ni de syntagmes croisés, ni de constituants récursifs*. L'intérêt de son utilisation comme élément de référence s'en est trouvé affaibli, en ce qui concerne nos besoins. Malgré un certain nombre de recherches, force nous a été de constater qu'il n'existait aucun corpus français disponible, renseigné des informations nécessaires à notre approche, à savoir la fonction syntaxique précise des constituants des phrases et leur position dans l'arbre syntaxique.

5.1.5. *Notre solution*

La seule alternative que nous ayons envisagée est de sélectionner minutieusement les corpus de tests, en préférant ceux qui sont les mieux analysés par SYGFRAN, puis de les transmettre à l'auteur de l'analyseur afin qu'il améliore les règles de SYGFRAN spécifiquement pour que ces textes soient parfaitement analysés. Ce processus demande un temps considérable car les règles sont complexes, nombreuses (SYGFRAN est composé de plus de 12 000 règles) et inter-dépendantes. Ainsi chaque modification ou ajout de règle risque de casser certaines analyses qui s'effectuaient correctement avant ces modifications. L'évaluation s'en trouve donc extrêmement ralentie. Néanmoins, elle est très présente à notre esprit. Contrairement aux applications pour lesquelles les résultats dépendent des mots, de leur fréquence ou de leur distribution, comme les campagnes d'évaluation TREC, les applications où les résultats dépendent de la structure sont plus difficiles à évaluer en raison des difficultés citées ci-dessus. Cela signifie que l'évaluation ne peut se faire qu'avec du temps, celui qui est nécessaire pour l'obtention d'un nombre de données non négligeable.

5.2. *Compression de phrases ou de segments ?*

La compression de phrases peut ne pas suffire à produire un résumé d'une taille convenable dans la plupart des cas. 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 réduisant davantage la taille du document est nécessaire.

Dans cette optique, nous cherchons à étendre la granularité des éléments supprimés et nous nous intéressons au résumé par suppression de segments textuels selon leur fonction en discours. Nous estimons que les fonctions discursives utiles au résumé incluent celles d'exemplification (Schiffrin *et al.*, 2003), de paraphrasage (Longacre, 1996) et d'explication car ces parties ont pour but de faciliter la compréhension du lecteur et leur suppression ne cause 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 et les marqueurs lexicaux.

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 dégrade ou améliore 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, en prétraitement.

6. Conclusion

Bien que le problème du résumé automatique ait déjà été abordé par de nombreux scientifiques depuis presque 50 ans (Luhn, 1958), l'approche que nous avons adoptée présente des éléments innovants. Les approches actuelles du résumé automatique utilisent des informations telles que 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, couplé aux règles de SYGFRAN, est l'un de ces outils. En outre, il ajoute à l'analyse en constituants de nombreuses informations concernant les relations entre constituants, nommées *dépendances*, ce que peu d'autres analyseurs proposent.

Notre approche a débuté par une étude sur l'importance des constituants dans une phrase, en s'appuyant en cela sur une démarche fortement qualitative et linguistique (par opposition à une démarche quantitative statistique). 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 % du texte de test, tout en conservant une très bonne cohérence grammaticale.

De cette étude de faisabilité, nous avons conclu que notre compression :

– est faisable si l'on a accès à un analyseur produisant des arbres syntaxiques et capable de donner des informations sur les dépendances ;

– est limitée à la granularité intra-phrastique, elle peut être utile dans des cas d'application précis, comme la réduction d'un texte narratif ;

– est étendue et confortée, elle peut participer activement à un processus plus large de résumé automatique, moyennant la vérification de ses conditions de validité.

Ce processus, vers lequel nous nous orientons, a été cité. Il se fonde sur la suppression de segments textuels ayant des fonctions discursives, soit conservatrices du thème comme les paraphrases et certaines explications, soit non conservatrices comme les exemples et certaines autres explications. 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.

Remerciements

Les auteurs de cet article remercient très vivement Augusta Mela, maître de conférences en sciences du langage, de sa contribution sur le plan linguistique, et les relecteurs de TSI pour leur travail critique de fond, qui a permis la consolidation de l'argumentation scientifique de cet article et un meilleur positionnement dans la défense de ses résultats.

7. Bibliographie

Alemany L. A., Fort M. F., « Integrating Cohesion and Coherence for Automatic Summarization », *EACL03*, Budapest, Hungary, avril, 2003.

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.

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.

Baldwin B., Morton T., « Dynamic coreference-based summarization », *in Proceedings of EMNLP-3 Conference*, 1998.

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

Boguraev B. K., Neff M. S., « Lexical Cohesion, Discourse Segmentation and Document Summarization », *RIAO-2000*, Paris, avril, 2000.

Brunet-Manquat F., *Outils génériques et robustes pour l'analyse de dépendances*, thèse de doctorat de l'Université Joseph Fourrier, Grenoble, 2004.

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

Chauché 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, p. 45-55, 2003.

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

Chomsky, *Some Concepts and Consequencesof the Theory of Government and binding*, Linguistic Inquiry monograph n°6, MIT Press, Cambridge, Mass., 1982.

Collins M., « Three generative lexicalized models for statistical parsing », *In proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, Madrid, Spain, p. 16-23, 1997.

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.

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, p. 391-407, 1990.

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.

Erkan G., Radev D. R., « LexRank : Graph-based Centrality as Salience in Text Summarization », *Journal of Artificial Intelligence Research (JAIR)*, 2004.

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.

Gendner V., Vilnat A., *Les annotations syntaxiques de référence PEAS*. mars, 2004, `http://www.limsi.fr/Recherche/CORVAL/easy/`.

Goldstein J., Mittal V., Carbonell J., Kantrowitz M., « Multi-document summarization by sentence extraction », *in Hahn et al.[15]*, p. 40-48, 2000.

Grefenstette G., « Producing intelligent telegraphic text reduction to provide audio scanning service for the blind », *In AAAI symposium on Intelligent Text Summarisation*, Menlo Park, California, p. 111-117, 1998.

Grevisse M., *le Bon Usage – Grammaire française*, édition refondue par André Goosse, DeBoeck-Duculot, Paris – Louvain-la-Neuve, 13e édition, ISBN 2-8011-1045-0, 1993-1997.

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, p. 342-348, août, 2002.

Ishikawa K., Ando S., 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.

Jing H., « Sentence Reduction for Automatic Text Summarization », *in Proceedings of the 6th Conference on Applied Natural Language Processing*, p. 310-315, 2000.

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*, p. 178-185, 2000.

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*, p. 68-73, 1995.

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, p. 703-710, 2000.

Knight K., Marcu D., « Summarization beyond sentence extraction : a probabilistic approach to sentence compression », *Artificial Intelligence archive*, vol. 139(1), p. 91-107, juillet, 2002.

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.

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

Longacre R. E., *The Grammar of Discourse*, Blackwell Publishing, 1996. ISBN 0-306-45235-9.

Luhn H., The automatic creation of literature abstracts., Journal of research and development, IBM, 1958.

Mani I., « Narrative Summarization », *in Proceedings of TAL, Résmé automatique de textes*, vol. 45/1, p. 15-38, 2004.

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, p. 243-281, 1988.

Marcu D., « Improving summarization through rhetorical parsing tuning », *in Proceedings of the COLING- ACL Workshop on Very Large Corpora*, Montreal, Canada, 1998.

Mauffrey A., Cohen I., *La grammaire française*, Hachette Education, 3ème édition, 1995.

McCord M., « English Slot Grammar », *IBM*, 1990.

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.

Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. J., « Introduction to wordnet : an on-line lexical database », *International Journal of Lexicography*, vol. 3(4), p. 235-244, 1990.

Minel J.-L., « Le résumé automatique de textes : solutions et perspectives », *in Proceedings of TAL, Résumé automatique de textes*, vol. 45/1, p. 7-13, 2004.

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 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, p. 344-348, 1994.

Quinlan J., « C4.5 : Programs for Machine Learning », *Morgan Kaufmann*, San Mateo, CA, 1993.

Radev D. R., Jing H., Styś M., Tam D., « Centroid-based summarization of multiple documents », *DUC 2003*, p. 919-938, décembre, 2004.

Radev D. R., McKeown K., « Generating Natural Language Summaries from Multiple On-Line Sources », *Computational Linguistics*, vol. 24, n° 3, p. 469-500, 1998.

Salton G., Yang C., « On the specification of term values in automatic indexing », *Journal of Documentation 29*, p. 351-372, avril, 1973.

Schiffrin D., Tannen D., Hamilton H. E., *The handbook of Discourse Analysis*, Blackwell Publishing, 2003. ISBN 0-631-20596-9.

Tomassone R., « A propos des "compléments circonstanciels" », *Les revues pédagogiques de la Mission Laïque Française*, p. 43-59, novembre, 2001.

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

Wagner R., Pinchon J., *Grammaire du Français classique et moderne*, Hachette Université, Paris, 1962.

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.

Mehdi Yousfi-Monod *est actuellement allocataire de recherche à l'Université de Montpellier II et prépare une thèse au sein du LIRMM (Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier). Ses travaux portent sur l'extraction et l'utilisation des informations syntaxiques des phrases (fonction syntaxique, rôle syntaxique, position dans l'arbre syntaxique, dépendances entre constituants) pour la compression de phrases comme tâche réalisée au sein du processus de résumé automatique.*

Violaine Prince *est professeur à l'Université Montpellier II, et dirige l'équipe « traitement algorithmique du langage » du LIRMM-CNRS. Dans son travail autour des applications de l'analyse syntaxique du français, elle s'intéresse aux applications en TAL de la notion linguistique d'effacement (qui se traduit ici par l'élagage sélectif de l'arbre syntaxique en vue de la contraction) et de la théorie de la divergence en traduction automatique. Elle développe un traducteur français-anglais, fondé sur l'analyseur SYGFRAN, et qui fonctionne par transformations successives des arbres syntaxiques entre langue source et langue cible.*

Article reçu le 03 janvier 2005
Version révisée le 15 septembre 2005