Simplification syntaxique de phrases pour le français

Laetitia Brouwers[1, 2] Delphine Bernhard[1, 3]
Anne-Laure Ligozat[1, 4] Thomas François[2, 5]
(1) LIMSI-CNRS, 91403 Orsay, France
(2) Université catholique de Louvain, Belgique
(3) LiLPa, Université de Strasbourg, France
(4) ENSIIE, Evry, France
(5) University of Pennsylvania, USA

RÉSUMÉ

Cet article présente une méthode de simplification syntaxique de textes français. La simplification syntaxique a pour but de rendre des textes plus abordables en simplifiant les éléments qui posent problème à la lecture. La méthode mise en place à cette fin s'appuie tout d'abord sur une étude de corpus visant à étudier les phénomènes linguistiques impliqués dans la simplification de textes en français. Nous avons ainsi constitué un corpus parallèle à partir d'articles de Wikipédia et Vikidia, ce qui a permis d'établir une typologie de simplifications. Dans un second temps, nous avons implémenté un système qui opère des simplifications syntaxiques à partir de ces observations. Des règles de simplification ont été décrites afin de générer des phrases simplifiées. Un module sélectionne ensuite le meilleur ensemble de phrases. Enfin, nous avons mené une évaluation de notre système montrant qu'environ 80% des phrases générées sont correctes.

ABSTRACT

Syntactic Simplification for French Sentences

This paper presents a method for the syntactic simplification of French texts. Syntactic simplification aims at making texts easier to understand by simplifying the elements that hinder reading. It is based on a corpus study that aimed at investigating the linguistic phenomena involved in the manual simplification of French texts. We have first gathered a parallel corpus of articles from Wikipedia and Vikidia, that we used to establish a typology of simplifications. In a second step, we implemented a system that carries out syntactic simplifications based on these corpus observations. We described simplification rules in order to generate simplified sentences. A module subsequently selects the best subset of sentences. The evaluation of our system shows that about 80% of the sentences produced by our system are accurate.

MOTS-CLÉS : simplification automatique, lisibilité, analyse syntaxique.

KEYWORDS: automatic simplification, readability, syntactic analysis.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 211–224,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Dans la majorité de nos activités quotidiennes, la capacité de lire rapidement et efficacement constitue un atout certain, voire un pré-requis. Willms (2003) souligne ainsi une corrélation entre ces compétences et le statut socio-économique des individus. Pourtant, une tranche non négligeable de la population n'est pas capable de traiter efficacement les données textuelles auxquelles ils sont confrontés. Richard *et al.* (1993) rapportent une expérience où, sur 92 demandes d'allocation de chômage remplies par des personnes avec un faible niveau d'éducation, pas moins de la moitié des informations requises (dont certaines étaient cruciales pour le traitement de la demande) manquaient, notamment à cause de problème de compréhension. Dans un contexte légèrement différent, à savoir la pharmacologie, Patel *et al.* (2002) parviennent à un constat similaire : la plupart de leurs sujets ont rencontré des problèmes importants dans la compréhension des différentes étapes à réaliser pour la bonne administration du médicament testé.

Ces problèmes de compréhension s'expliquent souvent par une trop grande complexité des textes, en particulier au niveau du lexique et de la syntaxe. Ces deux facteurs sont connus comme étant des causes importantes des difficultés de lecture (Chall et Dale, 1995), en particulier chez les jeunes enfants, les apprenants d'une langue étrangère ou les personnes présentant des déficiences intellectuelles.

Dès lors, la simplification automatique de textes apparaît comme un moyen susceptible d'aider ces personnes à accéder plus facilement au contenu des documents écrits auxquels ils sont confrontés. Il s'agit d'un domaine du traitement automatique des langues (TAL) visant à rendre des textes plus abordables tout en garantissant l'intégrité de leur contenu et en veillant à en respecter la structure. Dès lors, il faut déterminer d'une part quelles informations sont secondaires afin de les supprimer et de rendre les informations primordiales plus visibles et d'autre part quelles sont les constructions syntaxiques qui peuvent poser problème pour les simplifier.

Parmi les premiers efforts en ce sens, citons (Carroll *et al.*, 1999) et (Inui *et al.*, 2003), qui ont proposé des outils pour produire des textes plus abordables pour les personnes atteintes d'un handicap langagier tel que l'aphasie ou la surdité. Cependant, l'aide à la lecture ne s'adresse pas qu'aux lecteurs présentant des handicaps, mais aussi à ceux qui apprennent une langue (première ou seconde). Ainsi, Belder et Moens (2010) se sont intéressés à la simplification pour des enfants de langue maternelle anglaise, tandis que Siddharthan (2006), Petersen et Ostendorf (2007) et Medero et Ostendorf (2011) ont étudié la simplification pour les apprenants d'une langue seconde. La plupart de ces travaux concernent la langue anglaise, à l'exception de (Inui *et al.*, 2003) qui traitent également le japonais.

Parallèlement, la simplification automatique a également été utilisée comme un pré-traitement visant à augmenter l'efficacité d'opérations postérieures effectuées sur des textes. Les premiers, Chandrasekar *et al.* (1996) ont considéré que les phrases longues et complexes constituaient un obstacle pour l'analyse syntaxique ou la traduction automatique et que leur simplification préalable pouvait conduire à de meilleures analyses. Plus récemment, Heilman et Smith (2010) ont montré, quant à eux, qu'un texte simplifié produit de meilleurs résultats dans un contexte de génération automatique de questions. Du coté du biomédical, Lin et Wilbur (2007) et Jonnalagadda *et al.* (2009) ont optimisé l'extraction de données en simplifiant les textes lors d'un prétraitement.

La majorité des méthodes de simplification syntaxique proposées reposent sur un ensemble de règles de transformation définies manuellement pour être appliquées aux phrases. La simplification semble toutefois naturellement se prêter à l'utilisation de méthodes issues de la traduction automatique ou de l'apprentissage automatique, dont les modèles sont construits à partir de corpus comparables de textes complexes et simplifiés (Zhu *et al.*, 2010; Specia, 2010; Woodsend et Lapata, 2011). Les données utilisées dans ce cas sont notamment issues de Wikipédia en anglais et de Simple English Wikipedia, destinée aux enfants et aux locuteurs non natifs. L'encyclopédie Simple English Wikipedia compte à ce jour plus de 75 000 articles.

Il existe des projets comparables pour le français, Vikidia (voir Section 2.1) et Wikimini, mais ils ne sont pas aussi fournis que leur homologue anglophone. Par ailleurs, les différentes versions d'un article de Wikipédia ne sont pas strictement parallèles, ce qui complique encore l'apprentissage automatique. La méthode proposée dans cet article repose donc sur un ensemble de règles de simplification automatique qui ont été définies manuellement (voir Section 2.3), après étude de corpus. Nous utilisons la technique de la sur-génération, qui consiste à produire dans un premier temps un nombre important de simplifications possibles, avant de procéder à une sélection optimale des meilleures simplifications produites, à l'aide de la programmation linéaire en nombre entiers (PLNE, en anglais *Integer Linear Programming – ILP*). La PLNE permet de définir des contraintes qui régissent le choix du résultat fourni par l'outil de simplification automatique. Cette méthode a notamment été appliquée à la simplification de textes en anglais par (Woodsend et Lapata, 2011), (Belder et Moens, 2010), ainsi que par (Gillick et Favre, 2009) pour le résumé automatique.

Les apports de cet article sont les suivants : l'étude des procédés de simplification en français, et notamment la constitution d'un corpus de phrases parallèles, et une typologie des simplifications ; l'utilisation de critères originaux de sélection des phrases, tels que la liste orthographique de base de Nina Catach ou les mots-clés d'un texte. Nous présenterons tout d'abord le processus de constitution du corpus (Section 2.1), puis la typologie des simplifications observées (Section 2.2). Nous détaillerons ensuite le fonctionnement du système mis en œuvre, qui procède en deux temps : une surgénération de phrases simplifiées (Section 2.3.1), et une sélection des phrases correspondant à des critères de lisibilité (Section 2.3.2). Enfin, nous évaluerons cette simplification du point de vue de la correction des phrases générées, et analyserons les causes d'erreurs (Section 3).

2 Méthodologie

2.1 Présentation du corpus

Pour établir une typologie des règles de simplification, une étude sur corpus a été réalisée. Puisqu'il s'agit de déterminer les stratégies utilisées pour passer d'une phrase complexe à une phrase simplifiée, un corpus de phrases parallèles a été construit à partir d'articles des encyclopédies en ligne Wikipédia [1] et Vikidia [2]. Cette dernière est destinée aux jeunes de huit à treize ans et rassemble des articles plus accessibles, tant au niveau de la langue que du contenu. Afin de constituer ce corpus, nous sommes partis des articles de Vikidia et avons utilisé l'API MediaWiki

1. http://fr.wikipedia.org
2. http://fr.vikidia.org

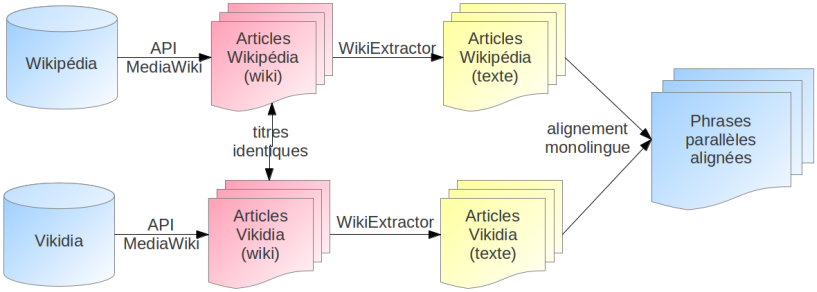

FIGURE 1 – Constitution du corpus de phrases parallèles

pour récupérer les articles de Wikipédia et Vikidia de mêmes titres. Le programme WikiExtractor [3] a ensuite été appliqué à ces articles afin d'en extraire les textes bruts (c'est-à-dire sans la syntaxe wiki). Le corpus ainsi constitué comprend 13 638 fichiers (dont 7 460 de Vikidia et 6 178 de Wikipédia, certains articles de Vikidia n'ayant pas d'équivalent direct dans Wikipédia).

Ces articles ont ensuite été analysés afin de repérer des phrases parallèles (phrase de Wikipédia ayant un équivalent simplifié dans Vikidia). Cet alignement a été effectué en partie manuellement et en partie automatiquement grâce à l'algorithme d'alignement monolingue décrit dans (Nelken et Shieber, 2006), qui se fonde sur une similarité cosinus entre phrases, avec un *tf.idf* adapté pour la pondération des mots. Ce programme fournit en sortie des alignements entre phrases, avec un score de confiance associé. La figure 1 résume le processus de constitution de ce corpus.

Parmi ces fichiers, vingt articles ou extraits d'articles de Wikipédia et leur équivalent dans Vikidia ont été sélectionnés, ce qui nous donne respectivement 72 phrases et 80 phrases. Les extraits suivants - correspondant à l'entrée «archipel» - ont par exemple été sélectionnés :

> (1a) Wikipédia : *Un archipel est un ensemble d'îles relativement proches les unes des autres. Le terme «archipel» vient du grec ancien "Archipelagos", littéralement «mer principale» (de "archi" : «principal» et "pélagos" : «la haute mer»). En effet, ce mot désignait originellement la mer Égée, caractérisée par son grand nombre d'îles (les Cyclades, les Sporades, Salamine, Eubée, Samothrace, Lemnos, Samos, Lesbos, Chios, Rhodes, etc.)*.
> (1b) Vikidia : *Un archipel est un ensemble de plusieurs îles, proches les unes des autres. Le mot «archipel» vient du grec "archipelagos", qui signifie littéralement «mer principale» et désignait à l'origine la mer Égée, caractérisée par son grand nombre d'îles.*

Notons que les deux articles présentent les mêmes informations globalement, mais de manière différente. Il y a simplification lexicale, sémantique et syntaxique. En effet, dans Vikidia, il n'y a que deux phrases, qui contiennent l'essentiel de l'explication (information nécessaire) tandis que dans Wikipédia, trois phrases détaillent la signification et l'origine du terme de manière plus précise (informations secondaires, par exemple mises entre parenthèses).

3. `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

2.2 Typologie de simplifications

Les observations réalisées sur ce corpus ont permis d'établir une typologie articulée selon trois grands niveaux de transformations : lexical, sémantique et syntaxique. Dans les travaux réalisés, la simplification est communément considérée comme composée de deux catégories, lexicale et syntaxique (Carroll *et al.*, 1999; Inui *et al.*, 2003; Belder et Moens, 2010). Le domaine de la sémantique quant à lui n'est pas cité. Ces trois grands niveaux peuvent être à leur tour divisés en sous-catégories, comme le montre la table 1.

Lexique	**Sémantique**	**Syntaxe**
Synonyme ou hyperonyme Traduction	Réorganisation Suppression Ajout	Temps Suppression Modification Division Regroupement

TABLE 1 – Typologie

En ce qui concerne le lexique, deux phénomènes sont observés. D'une part, les termes considérés comme difficiles sont remplacés par un synonyme ou un hyperonyme. Dans l'exemple (1), *terme* a été remplacé par *mot* qui est plus courant. D'autre part, les concepts utilisés dans leur langue d'origine dans Wikipédia sont traduits en français dans Vikidia.

Au niveau sémantique, les auteurs de Vikidia prêtent une attention particulière à l'organisation de l'information qui doit être claire et synthétique. Dans cette optique, il arrive que des propositions soient interverties, afin d'assurer une meilleure présentation de l'information. De plus, le contenu considéré comme secondaire à la compréhension est supprimé tandis que des explications ou des exemples sont ajoutés pour plus de clarté. Ainsi, dans l'exemple (1), la décomposition de la signification du mot *archipel* est explicitée dans Wikipédia, mais pas dans Vikidia.

Enfin, du point de vue syntaxique, qui nous intéresse prioritairement ici, cinq types de changements sont observés : les modifications de temps, la suppression, la modification, la division et le regroupement. Les deux derniers types peuvent être envisagés ensemble dans la mesure où ce sont deux phénomènes opposés. Cette classification peut se rapprocher de celle de (Medero et Ostendorf, 2011) qui reprend trois catégories - la division, la suppression et l'extension - ou de (Zhu *et al.*, 2010) (composée de la division, la suppression, la réorganisation et la substitution).

– Tout d'abord, les temps utilisés dans Vikidia sont plus quotidiens et moins littéraires que ceux utilisés dans Wikipédia. Ainsi, le présent et le passé composé sont préférés au passé simple.
– Ensuite, les informations secondaires ou redondantes, telles que certains compléments circonstanciels, qui sont en général considérées comme supprimables au niveau syntaxique, ne sont pas reprises dans les articles de Vikidia. Dans l'exemple (1), l'adverbe *relativement* qui précédait *proches les unes des autres* a ainsi été supprimé dans Vikidia. L'adverbe n'ajoutait effectivement rien au niveau informationnel.
– De plus, si certaines structures plus complexes ne sont pas supprimées, elles sont alors déplacées ou modifiées pour plus de clarté. Dans Vikidia, par exemple, une construction affirmative est préférée à une forme négative :

> (2a) Wikipédia : *Les personnes qui ont voté blanc ou nul ne sont généralement pas considérées comme abstentionnistes mais le résultat est identique : leur choix n'est pas*

pris en compte.
(2b) Vikidia : *L'abstention est différente du vote blanc et du vote nul.*

– Finalement, les auteurs choisissent parfois de diviser des phrases longues ou à l'inverse de réunir plusieurs phrases en une seule. Dans l'exemple (1), les deux dernières phrases ont été regroupées dans Vikidia, car elles ont été simplifiées et sont dès lors devenues beaucoup plus courtes. Il faut d'emblée préciser que le regroupement d'éléments est beaucoup moins utilisé que la division de phrases. Pour scinder une phrase, les auteurs prennent par exemple une proposition secondaire (telle qu'une relative) qu'ils transforment en phrase indépendante.

Parmi les changements observés, certains d'entre eux sont difficilement implémentables. C'est le cas lorsqu'une modification nécessite de recourir à la sémantique, c'est-à-dire qu'il n'est possible de repérer les structures à modifier que par le sens. Il est difficile d'appliquer ce type de stratégies de manière automatique. Par exemple, il est parfois possible de supprimer les éléments qui se rapportent au nom, alors que d'autre fois, ils sont indispensables, sans que cela ne soit marqué typographiquement ou grammaticalement dans la phrase.

D'autres changements syntaxiques doivent s'accompagner de transformations lexicales, difficilement généralisables. Par exemple, la modification d'une phrase négative en une phrase affirmative nécessite de trouver un verbe dont la forme affirmative recouvre le sens de la construction négative à remplacer.

Il y a également des changements qui sont effectués de manière isolée et non systématisable. Ils relèvent plutôt d'un traitement manuel que d'un traitement automatique d'un texte, dans le sens où chaque cas est différent (même s'il s'inscrit dans une règle plus globale). De plus, ils font généralement appel à des informations sémantiques ou lexicales et pas simplement syntaxiques. Il s'agit de changements complexes, qui sont utiles dans certains cas, mais ardus à détecter automatiquement.

Enfin, les changements syntaxiques qui ont un impact sur d'autres parties du texte ou qui concernent des éléments dépendants d'une autre structure demandent des modifications plus globales du texte. Par conséquent, ils sont également difficiles à traiter automatiquement. Ainsi, pour modifier le temps d'un verbe dans une phrase, il faut veiller à ce que la concordance des temps soit respectée dans l'entièreté du texte.

2.3 Système de simplification syntaxique

Nous avons utilisé cette typologie pour mettre en œuvre un système de simplification syntaxique pour le français. La simplification d'un texte y est effectuée en deux étapes : une étape de génération de toutes les simplifications possibles pour chaque phrase du texte, et une étape de sélection du meilleur ensemble de phrases simplifiées. L'architecture de ce système est présentée dans la figure 2.

Le module de surgénération s'appuie sur un ensemble de règles (au nombre de 19), utilisant des informations sur les caractéristiques (morpho-)syntaxiques des mots et sur les relations de dépendance présentes au sein d'une phrase. C'est pourquoi les textes de notre corpus ont été analysés par MElt [4] (Denis *et al.*, 2009) et Bonsai [5] (Candito *et al.*, 2010). Ces textes ont ainsi été

4. https://gforge.inria.fr/projects/lingwb
5. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

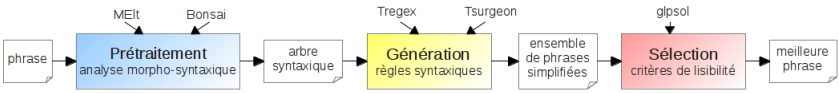

FIGURE 2 – Organisation du système de simplification syntaxique

représentés sous la forme d'arbres syntaxiques, lesquels contiennent un maximum de données utiles à l'application de règles de simplification. Ces dernières peuvent alors être appliquées de manière récursive, jusqu'à ce qu'il n'y ait plus aucune structure à simplifier dans chacune des phrases des textes. Il faut ajouter que toutes les phrases créées à chaque application d'une règle sont enregistrées, produisant un ensemble de variantes. Par la suite, le meilleur ensemble de phrases sera retenu via un modèle de programmation linéaire, en fonction d'une série de critères détaillés par la suite.

2.3.1 Génération de phrases simplifiées

Les règles de simplification syntaxique qui composent notre programme sont respectivement des règles de suppression (12 règles), de modification (3 règles) et de division (4 règles). Notons que, par rapport à la typologie établie, deux types de règles n'ont pas été mises en place. D'une part, les stratégies de regroupement de plusieurs phrases en une n'ont pas été observées de manière assez systématique dans le corpus d'étude. Il est dès lors difficile d'en retirer une règle automatisable. De plus, les règles de regroupement pourraient entrer en conflit avec les règles de suppression, puisqu'elles ont des buts opposés. D'autre part, en ce qui concerne les aspects temporels, nous avions noté que certains temps étaient plus utilisés que d'autres dans l'encyclopédie pour les jeunes, Vikidia. Toutefois, cette stratégie n'a pas été implémentée car elle demandait des changements trop globaux, pouvant toucher au texte entier. En effet, lorsqu'un verbe au passé simple est remplacé par un verbe au présent, il faut veiller à ce que la concordance des temps soit toujours respectée partout, ce qui demande d'examiner tout le texte, ou du moins le paragraphe qui contient la forme verbale modifiée. On risque alors de détruire la cohérence du texte et d'en altérer la qualité.

Pour appliquer ces 19 règles, il convient tout d'abord de repérer les structures concernées par de possibles changements à l'aide d'expressions régulières et grâce à `Tregex`[6] (Levy et Andrew, 2006) qui gère le repérage d'éléments et de relations dans un arbre. Dans un deuxième temps, une série d'opérations sont effectuées par le biais de `Tsurgeon` qui permet de modifier des arbres syntaxiques. Par exemple, pour supprimer une coordonnée introduite par *soit*, il faut repérer une proposition coordonnée, étiquetée `COORD`, qui domine la conjonction de coordination *soit* et lui donner un nom comme `Pcoord`. Ensuite, l'opération `Tsurgeon delete` doit être appliquée à l'ensemble repris sous `Pcoord` :

Repérage (`Tregex`) : `COORD=Pcoord < (CC < /soit/)`
Opération (`Tsurgeon`) : `delete Pcoord`

6. `http://nlp.stanford.edu/software/tregex.shtml`

Cette règle s'appliquerait par exemple à la phrase suivante :

> (3a) Phrase d'origine : *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville, soit 20 % de la population totale du pays.*
> (3b) Phrase après application de la règle : *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville.*

Les opérations varient en fonction du type de règle appliquée :

1. Pour les règles de suppression, il suffit de supprimer tous les éléments concernés (via l'opération `Tsurgeon delete`). Les éléments concernés par les règles de suppression sont les compléments circonstanciels, les ensembles entre parenthèses, une partie des propositions subordonnées, les propositions entre virgules ou introduites par un terme tel que *comme*, *voire* ou *soit*, les adverbes et les compléments d'agent.

2. Pour les règles de modification, il s'agit de combiner plusieurs opérations : la suppression de certains termes (opération `Tsurgeon delete`), le déplacement d'éléments (opération `Tsurgeon move`) et l'ajout d'étiquettes (opération `Tsurgeon insert`) qui signalent un traitement éventuel par la suite. En effet, certaines règles demandent que les formes verbales se conjuguent à un autre temps, un autre mode, etc. Dans ce cas, des étiquettes sont ajoutées autour du verbe pour indiquer qu'il doit être modifié. Il sera, dans un traitement postérieur, conjugué à la forme voulue grâce au système de conjugaison le `Verbiste`[7]. Ainsi, pour passer d'une structure passive à une structure active, il faut modifier le mode, mais aussi parfois la personne, pour que le verbe s'accorde correctement avec le complément d'agent devenu sujet. Trois règles de modification ont été mises en place : le déplacement à l'initiale des compléments circonstanciels, le passage à l'actif des formes passives et la transformation d'une clivée en non clivée.

3. Pour les règles de division, le processus se déroule en deux étapes. La proposition secondaire est d'abord supprimée et la nouvelle phrase enregistrée telle quelle. Ensuite, la phrase d'origine est reprise et la proposition principale est cette fois supprimée avant que la proposition secondaire soit transformée de manière à devenir indépendante. En général, il faut veiller à modifier la forme verbale de cette proposition secondaire pour qu'elle puisse fonctionner comme un verbe principal. Par ailleurs, le pronom qui régit la proposition doit être remplacé par son antécédent et le sujet doit être inséré lorsqu'il est manquant. Par exemple, pour transformer une relative en une proposition indépendante, le pronom relatif doit être remplacé par son antécédent et il est important de tenir compte de la fonction du pronom pour savoir où insérer l'antécédent. Signalons que les phrases sont scindées quand elles contiennent des propositions secondaires introduites par deux points, des coordonnées, des participiales ou des relatives.

Ces règles de simplification sont appliquées de manière récursive à une phrase jusqu'à ce que toutes les variantes possibles aient été générées. Plusieurs résultats sont donc régulièrement obtenus pour une même phrase. Dès lors, il convient de déterminer la phrase, parmi toutes celles produites, qui est la plus appropriée pour remplacer celle d'origine. Ce processus est décrit à la section suivante.

7. Le programme est disponible à l'adresse `http://sarrazip.com/dev/verbiste.html` sous licence GNU (page consultée le 6 novembre 2011). Il a été créé par Pierre Sarrazin.

2.3.2 Sélection de phrases simplifiées

Étant donné un ensemble de phrases simplifiées possible pour un texte, notre objectif est de sélectionner le meilleur sous-ensemble de phrases simplifiées, c'est-à-dire celui qui maximise une mesure de lisibilité. Cette mesure de lisibilité se traduit par différents critères. Pour résoudre ce genre de problèmes, la programmation linéaire en nombres entiers constitue une technique appropriée.

Dans notre cas, quatre critères ont été pris en compte pour choisir la phrase adéquate : la longueur de la phrase, la longueur des mots, la familiarité du vocabulaire et la présence de termes-clés, c'est-à-dire récurrents dans le texte. La longueur de la phrase est exprimée en nombre de mots tandis que la longueur des mots est donnée en nombre de caractères. En ce qui concerne la familiarité des mots, la liste de Catach [8] (Catach, 1985) a été utilisée pour calculer le poids de chaque terme. Il s'agit d'une liste des 3000 mots les plus fréquents, dont il convient d'enseigner l'orthographe en priorité aux élèves de primaire. Les termes-clés ont été définis, quant à eux, comme les mots qui apparaissaient deux fois ou plus dans un texte.

Ces critères sont combinés grâce à la formule suivante au sein du module de programmation linéaire [9] :

$$\begin{array}{ll} \text{Il s'agit alors de maximiser :} & h_w + h_s + h_a + h_c \\ \text{Où :} & h_w = \text{wps} \times \sum_i s_i - \sum_i l_i^w s_i \\ & h_s = \text{cpw} \times \sum_i l_i^w s_i - \sum_i l_i^c s_i \\ & h_a = \text{aps} \times \sum_i s_i - \sum_i l_i^a s_i \\ & h_c = \sum_j w_j c_j \end{array} \quad (1)$$

Nous avons défini les paramètres et variables suivants pour la formulation du problème :
– wps : le nombre moyen de mots par phrase souhaité
– cpw : le nombre moyen de caractères par mot souhaité
– aps : le nombre moyen de mots absents de la liste de Catach souhaité
– s_i un indicateur de la présence de la phrase i dans la simplification de texte finale
– c_j un indicateur de la présence du mot-clé j dans la simplification de texte finale
– l_i^w la longueur en mots de la phrase i
– l_i^c la longueur en caractères de la phrase i
– l_i^a le nombre de mots absents de la phrase i
– w_j le nombre d'occurrences du mot-clé j

`wps`, `cpw` et `aps` sont des paramètres constants dont les valeurs ont été fixées respectivement à 10, à 5 et à 2 pour cette étude. Toutefois, il s'agit de paramètres susceptibles de varier en fonction du contexte d'utilisation et du public cible, puisqu'ils déterminent directement le niveau de difficulté des phrases simplifiées retenues.

Pour illustrer ce processus, prenons le texte de départ pour l'article de Wikipédia intitulé *Abel*. Il comprenait 25 phrases, à partir desquelles 67 phrases simplifiées ont été produites. Parmi le texte simplifié, nous observons que ce sont les phrases 3 de l'exemple (4b) qui remplacent la phrase du texte original (exemple (4a)) :

> (4a) Phrase d'origine (Phrase 1) : *Caïn, l'aîné, cultive la terre et Abel (étymologie : de l'hébreu " souffle ", " vapeur ", " existence précaire ") garde le troupeau.*

8. Elle est notamment disponible sur le site `http://www.ia93.ac-creteil.fr/spip/spip.php?article2900`.
9. Le module repose sur `glpk` qui est disponible à l'adresse suivante : `http://www.gnu.org/software/glpk/`

(4b) Simplifications possibles :
Phrase 2 : *Caïn, l'aîné, cultive la terre et Abel garde le troupeau.*
Phrases 3 : *Caïn, l'aîné, cultive la terre. Abel garde le troupeau.*
Phrase 4 : *Caïn, l'aîné, cultive la terre.*
Phrase 5 : *Abel garde le troupeau.*
Phrases 6 : *Caïn, l'aîné, cultive la terre. Abel (étymologie : de l'hébreu " souffle ", " vapeur ", " existence précaire ") garde le troupeau.*
Phrase 7 : *Abel (étymologie : de l'hébreu " souffle ", " vapeur ", " existence précaire ") garde le troupeau.*
(4c) Simplification sélectionnée (Phrase 3) : *Caïn, l'aîné, cultive la terre. Abel garde le troupeau.*

Les valeurs des paramètres pour chaque phrase de l'exemple (4) sont données dans la table 2.

	Longueur de la phrase	**Longueur des mots**	**Familiarité des mots**	**Termes clés**
Valeurs souhaitées	10 mots	5 caractères	2 mots absents	
Phrase 1	19 mots	6,1 caractères	11 mots absents	5 termes
Phrase 2	11 mots	4,3 caractères	5 mots absents	5 termes
Phrases 3	5 mots	4,6 caractères	2 mots absents	5 termes
Phrase 4	6 mots	4,5 caractères	3 mots absents	3 termes
Phrase 5	4 mots	4,7 caractères	2 mots absents	2 termes
Phrases 6	9 mots	6,3 caractères	5 mots absents	5 termes
Phrase 7	12 mots	7,3 caractères	8 mots absents	2 termes

TABLE 2 – Valeurs des paramètres pour les phrases de l'exemple (4)

3 Évaluation

La simplification syntaxique implique des modifications importantes au sein de la phrase aussi bien au niveau du contenu que de la forme. C'est pourquoi il est important de vérifier que l'application d'une règle ne provoque pas des erreurs qui rendraient les phrases produites incompréhensibles ou agrammaticales. Une évaluation manuelle de notre système de génération de phrases simplifiées a donc été réalisée dans ce but. Elle repose sur un nouveau corpus composé de neuf articles de Wikipédia, c'est-à-dire de 202 phrases. Les résultats obtenus sont détaillés à la table 3 et dans la Section 3.1. Nous y détectons deux grands types d'erreurs, à savoir les erreurs d'analyse (morpho-)syntaxique et les erreurs de simplification. Celles-ci sont discutées dans les Sections 3.2 et 3.3.

3.1 Données obtenues

Sur les 202 phrases qui composent le corpus d'évaluation, 113 d'entre elles (56%) ont subi une ou plusieurs simplifications. Ces 113 phrases auxquelles des règles ont pu être appliquées donnent lieu à 333 variantes susceptibles de comporter des erreurs. C'est effectivement le cas de 71 d'entre elles (21,32%). Parmi celles-ci, il faut distinguer d'emblée les erreurs dues au

programme de simplification et les erreurs provoquées par un pré-traitement (analyse morpho-syntaxique et syntaxique) inexact. Ainsi, parmi les 21,32% de phrases problématiques, il apparaît que 89% des erreurs proviennent de l'analyse (morpho-)syntaxique et seulement 11% d'entre elles sont effectivement dues au système mis en place. Par conséquent, à partir du corpus d'évaluation, seulement 2,4% des phrases produites par notre système posent problème en raison de l'application d'une règle de simplification. Parmi ces rares erreurs de simplification (notre corpus en compte 8), il faut enfin distinguer les erreurs de contenu (25%) et de forme (75%). Dans la suite de cette section, nous revenons plus en détail sur les deux types d'erreurs principales rencontrées : morpho-syntaxiques et de simplification.

<table>
<tr><td colspan="4">Phrases produites par le programme
333 phrases - 100%</td></tr>
<tr><td>Phrases correctes
262 phrases - 78,68%</td><td colspan="3">Phrases incorrectes
71 phrases - 21,32%</td></tr>
<tr><td></td><td>Erreurs dues au pré-traitement
63 phrases - 18,92%</td><td colspan="2">Erreurs dues au simplificateur
8 phrases - 2,4%</td></tr>
<tr><td></td><td></td><td>Syntaxe
6 phrases
1,8%</td><td>Sens
2 phrases
0,6%</td></tr>
</table>

TABLE 3 – Évaluation des règles de simplification

3.2 Erreurs d'analyse (morpho-)syntaxique

La phase de pré-traitement consiste à étiqueter, annoter et structurer des phrases. Dès lors, il peut y avoir des erreurs dans les étiquettes attribuées, les relations identifiées, les regroupements d'éléments et les délimitations de phrases et de parenthèses.

Les erreurs d'étiquette sont les plus fréquentes et concernent des entités nommées, des cas ambigus ou des expressions figées. Par exemple, les mots *ainsi que* peuvent poser problème puisqu'il peut s'agir d'un connecteur ou des termes *ainsi* et *que*. L'analyseur ne parvient pas toujours à différencier les deux cas, ce qui peut provoquer des erreurs lors de la suppression des coordonnées (si les mots *ainsi que* sont identifiés comme un connecteur à tort). La phrase suivante en est un exemple :

> (5) *Les mélodies sont accrocheuses et les arrangements très soignés ; c'est ainsi que "Mamma Mia" et "Fernando" (malgré quelques erreurs de grammaire anglaise) occupèrent la première place des palmarès mondiaux dans le premier semestre de cette même année.*

Puisque *ainsi que* est considéré comme un connecteur et non comme l'adverbe *ainsi* suivi du deuxième terme de la clivée *que*, la règle de suppression des coordonnées produit la phrase suivante qui est agrammaticale : *C'est*.

Au-delà des problèmes d'étiquette, les relations de dépendance posent aussi fréquemment des difficultés à l'analyseur syntaxique. En effet, il est difficile de déterminer où s'arrête un groupe ou une proposition, quels éléments le composent ou de quel élément il dépend, particulièrement lorsque les constructions sont complexes et même emboitées. À nouveau, cela peut poser problème si une règle de simplification s'applique justement à un groupe mal analysé.

Les ponctuations constituent également des éléments difficiles à traiter pour l'analyseur. C'est pourquoi les phrases et les groupes entre parenthèses ne sont pas toujours convenablement délimités. De plus, l'analyseur ne distingue pas les points contenus dans les citations entre guillemets et ceux qui marquent une fin de phrase. Cela peut amener le programme, qui applique des règles, à diviser une phrase en plusieurs propositions de manière erronée. Les parenthèses, quant à elles, ne sont pas toujours marquées à un même niveau, ce qui détruit l'unité de l'ensemble. En effet, tous les composants de l'expression entre parenthèses ne sont pas rassemblés sous un même élément, ce qui signifie qu'une partie de l'expression peut être supprimée sans le reste et inversement.

3.3 Erreurs de simplification

À côté des erreurs issues du pré-traitement, certaines phrases, erronées aux niveaux sémantique et syntaxique, peuvent être produites à la suite de l'application des règles de simplification. Il s'agit évidemment là des erreurs les plus intéressantes, qui se répartissent en deux grandes catégories.

D'une part, les informations véhiculées par la phrase peuvent se trouver modifiées ou amputées. De fait, lors de la suppression de l'infinitive, il arrive qu'une partie du contenu de la phrase soit perdue. Ainsi, dans la phrase suivante, issue de l'article *abbé*, la proposition infinitive, qui explique le terme *abbé*, est supprimée :

(6a) *C'est aussi depuis le XVIIIe siècle le terme en usage pour désigner un clerc séculier ayant au moins reçu la tonsure.*
(6b) *C'est aussi depuis le XVIIIe siècle le terme en usage.*

Par ailleurs, lors de la suppression du complément d'agent, le sens d'une phrase peut être bouleversé par ce type de modification. Il en est ainsi pour la phrase de l'exemple (7b). En effet, le sens de la phrase originale (7a) était tout à fait différent :

(7a) *Ils ne sont pas caractérisés par leur profession comme dans la Bible : l'un pasteur, l'autre agriculteur.*
(7b) *Ils ne sont pas caractérisés : l'un pasteur, l'autre agriculteur.*

D'autre part, la structure de la phrase peut être modifiée de telle façon que la phrase devient syntaxiquement incorrecte. Trois règles de simplification sont concernées. Tout d'abord, les règles de suppression sont sujettes à ce genre de problème puisqu'il s'agit de supprimer une partie de la phrase, qui est normalement secondaire au bon fonctionnement de la phrase. Pourtant, il arrive que l'élément supprimé soit essentiel, comme dans le cas de la suppression du référent d'un pronom. La suppression de la subordonnée ou de l'infinitive peut provoquer ce type de désagrément. De son côté, la division de la phrase à partir de la relative produit un autre genre d'erreurs. Si le verbe est suivi d'un infinitif, le complément direct (l'antécédent du relatif qui est replacé dans la relative lors de la division) peut dépendre du verbe ou de l'infinitif. Le simplificateur n'en tient pas compte et estime dans tous les cas que le complément direct dépend du verbe et non de l'infinitif, parfois de manière erronée comme dans l'exemple (8) :

(8a) *Ils ont sur leurs religieux un droit de juridiction, une autorité qu'il leur est recommandé de n'exercer que par la voie de la patience et de la douceur.*
(8b) *Il leur est recommandé cette autorité de n'exercer que par la voie de la patience et de la douceur.*

4 Conclusion et perspectives

Cet article a décrit un système automatique de simplification syntaxique pour le français à destination des enfants en particulier. Celui-ci repose sur un ensemble de règles obtenues sur la base d'une étude de corpus, laquelle a aussi mené à l'élaboration d'une typologie des simplifications en français. Il serait aisé d'étendre notre typologie à d'autres publics sur base d'autres corpus adéquats. Notre démarche utilise également la technique de la sur-génération, qui permet de retenir le meilleur ensemble de simplifications en fonction de critères de lisibilité. Notons que parmi ceux employés, certains n'avaient pas été considérés précédemment et produisent des résultats intéressants. Enfin, il est apparu que les performances de notre système sont bonnes (environ 80% des phrases générées sont correctes), en particulier si l'on ne tient pas compte des erreurs dues aux outils de prétraitement.

Nous envisageons plusieurs perspectives d'amélioration pour notre système. Tout d'abord, la simplification syntaxique pourrait être complétée par une simplification lexicale, ainsi que cela est fait dans certaines études pour l'anglais (Woodsend et Lapata, 2011). Il s'agit en effet d'une autre source de problèmes pour certains lecteurs. Par ailleurs, notre analyse des erreurs a souligné la nécessité d'ajouter ou de répéter des termes lorsqu'une division de phrase est effectuée. Il serait dès lors utile de développer un outil qui gérerait les référents, afin d'améliorer la qualité du texte simplifié. Enfin, une dernière perspective d'amélioration consisterait à rendre le système de règles modulable en fonction d'un public cible. Cela demanderait d'évaluer la pertinence des différentes transformations et des critères de sélection des meilleures simplifications en fonction des publics visés. Cette perspective nécessiterait d'évaluer l'efficacité des règles au moyen de tests de compréhension portant sur des phrases originales et simplifiés.

Remerciements Nous remercions Antoine Sylvain pour sa participation à la constitution du corpus de textes issus de Wikipédia et Vikidia. Ces travaux ont reçu le soutien financier du projet DOXA du pôle de compétitivité CAP-DIGITAL.

Références

BELDER, J. D. et MOENS, M.-F. (2010). Text Simplification for Children. *In Proceedings of the Workshop on Accessible Search Systems, in conjunction with SIGIR 2010*.

CANDITO, M., NIVRE, J., DENIS, P. et ANGUIANO, E. (2010). Benchmarking of statistical dependency parsers for French. *In Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, pages 108–116. Association for Computational Linguistics.

CARROLL, J., MINNEN, G., PEARCE, D., CANNING, Y., DEVLIN, S. et TAIT, J. (1999). Simplifying Text for Language-Impaired Readers. *In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270.

CATACH, N. (1985). *Les listes orthographiques de base du français*. Nathan, Paris.

CHALL, J. et DALE, E. (1995). *Readability Revisited : The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

CHANDRASEKAR, R., DORAN, C. et SRINIVAS, B. (1996). Motivations and methods for text simplification. *In Proceedings of the 16th conference on Computational linguistics*, pages 1041–1044.

Denis, P, Sagot, B. *et al.* (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of PACLIC*.

Gillick, D. et Favre, B. (2009). A scalable global model for summarization. *In Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, ILP '09, pages 10–18, Stroudsburg, PA, USA.

Heilman, M. et Smith, N. A. (2010). Extracting Simplified Statements for Factual Question Generation. *In Proceedings of the 3rd Workshop on Question Generation*.

Inui, K., Fujita, A., Takahashi, T., Iida, R. et Iwakura, T. (2003). Text simplification for reading assistance : a project note. *In Proceedings of the second international workshop on Paraphrasing*, pages 9–16.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C. et Gonzalez, G. (2009). Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. *In Proceedings of NAACL-HLT 2009*.

Levy, R. et Andrew, G. (2006). Tregex and tsurgeon : tools for querying and manipulating tree data structures. *In Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234.

Lin, J. et Wilbur, W. J. (2007). Syntactic sentence compression in the biomedical domain : facilitating access to related articles. *Information Retrieval*, 10(4):393–414.

Medero, J. et Ostendorf, M. (2011). Identifying Targets for Syntactic Simplification. *In Proceedings of the SLaTE 2011 workshop*.

Nelken, R. et Shieber, S. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 161–168.

Patel, V., Branch, T. et Arocha, J. (2002). Errors in interpreting quantities as procedures : The case of pharmaceutical labels. *International journal of medical informatics*, 65(3):193–211.

Petersen, S. E. et Ostendorf, M. (2007). Text Simplification for Language Learners : A Corpus Analysis. *In Proceedings of Speech and Language Technology in Education (SLaTE2007)*, pages 69–72.

Richard, J., Barcenilla, J., Brie, B., Charmet, E., Clement, E. et Reynard, P. (1993). Le traitement de documents administratifs par des populations de bas niveau de formation. *Le Travail Humain*, 56(4):345–367.

Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109.

Specia, L. (2010). Translating from Complex to Simplified Sentences. *In Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language (Propor-2010).*, pages 30–39.

Willms, J. (2003). Literacy proficiency of youth : Evidence of converging socioeconomic gradients. *International Journal of Educational Research*, 39(3):247–252.

Woodsend, K. et Lapata, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK.

Zhu, Z., Bernhard, D. et Gurevych, I. (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. *In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China.