Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Simplification de textes : un état de l'art

Sofiane ELGUENDOUZE[1,2]
(1) LS2N, Nantes, France
(2) ESI, Alger, Algérie
fs_elguendouze@esi.dz

RÉSUMÉ

Cet article présente l'état de l'art en simplification de textes et ses deux grandes familles d'approches, à savoir les approches à base de règles et les approches statistiques. Nous présentons, en particulier, les récentes approches neuronales et les architectures mises en place ainsi que les méthodes d'évaluation des systèmes de simplification.

ABSTRACT

Text simplification (State of the art)

This paper presents the state of the art on text simplification, in particular the two main types of approaches, namely rule-based approaches and statistical approaches (or data-driven). We present, in particular, the recent neural approaches and the architectures implemented as well as the evaluation methods of text simplification systems.

MOTS-CLÉS : Simplification de textes, Apprentissage automatique, Apprentissage profond, Traduction automatique, Lexique, Syntaxe, Discours.

KEYWORDS: Text simplification, Machine learning, Deep learning, Machine translation, Lexicon, Syntax, Discourse.

1 Introduction

Les textes incorporent des constructions linguistiques complexes, ce qui peut entraîner des difficultés de lecture et/ou de compréhension chez les personnes avec des compétences linguistiques réduites comme les apprenants d'une langue non native ou les personnes ayant un trouble du langage telles que les autistes, les aphasiques, les dyslexiques etc.
La simplification de Texte (ST) est définie comme un processus permettant de détecter les phénomènes problématiques dans un texte, causant des difficultés de lecture et/ou de compréhension, et de procéder à l'adaptation de son contenu pour les résoudre. Cela peut être effectué par le remplacement des mots difficiles par des synonymes plus faciles à comprendre, le changement des temps de conjugaison difficiles par d'autres plus courants ou encore par la résolution d'anaphores. Tout cela a pour objectif de rendre les textes plus accessibles compréhensibles par une catégorie de lecteurs spécifiques.

Les premiers travaux de recherche sur la simplification de textes ont commencé dans les années 90, avec l'approche de simplification syntaxique proposée par (Chandrasekar *et al.*, 1996), qui visait à améliorer les performances des analyseurs syntaxiques en langage naturel. Des recherches ultérieures ont principalement porté sur la ST pour des catégories d'utilisateurs avec des déficiences intellectuelles spécifiques. Certains travaux se sont concentrés sur l'aphasie (Carroll *et al.*, 1998),

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 96–109.
HAL : hal-02786191.

la dyslexie (Quiniou & Daille, 2018), l'autisme (Štajner *et al.*, 2012), les personnes avec un faible niveau de lecture (Williams & Reiter, 2008), les personnes sourdes (Inui *et al.*, 2003), les apprenants d'une langue étrangère (Siddharthan & Katsos, 2010). L'intérêt porté à la simplification de texte a récemment augmenté grâce à la disponibilité des textes et l'explosion du web etc., et le nombre de langues sur lesquelles elle est appliquée ne cesse d'accroître : l'anglais (Siddharthan, 2006; Zhu *et al.*, 2010; Xu *et al.*, 2016), le français (Seretan, 2012; Brouwers *et al.*, 2012, 2014; Gala *et al.*, 2018), l'espagnol (Saggion *et al.*, 2015), le portugais (Aluísio & Gasperin, 2010) ...
L'objectif de ce papier est de mettre en avant les dernières techniques de simplification notamment les approches neuronales qui ont largement contribué à l'amélioration des systèmes de simplification. La suite du papier présentera une typologie des phénomènes problématiques nécessitant une simplification, les différents types approches et les mesures d'évaluation les plus utilisées actuellement.

2 Taxonomie des problèmes

La complexité des textes et les informations implicites contenues dans ceux-ci affectent une tranche considérable de la population qui souffre de difficultés de lecture et de compréhension. Cela montre qu'il existe une multitude de domaines pour lesquels la simplification de textes pourra être utilisée. Plusieurs phénomènes problématiques tels que les phrases longues, les phrases complexes ou encore les pronoms et leurs référents implicites, sont généralement difficiles à comprendre par le lecteur. Des constructions syntaxiques telles que les phrases qui ne respectent pas la forme canonique (sujet, verbe, complément) peuvent également être problématiques pour les personnes aphasiques ou autistes. Il en va de même pour le vocabulaire très difficile ou spécialisé et les mots rares qui peuvent être ambigus. Les apprenants d'une langue étrangère peuvent avoir un lexique très restreint et ne seraient donc pas en mesure de comprendre certaines constructions grammaticales complexes ou certains mots difficiles. La Table 1 présente une synthèse d'analyse de certaines typologies de phénomènes linguistiques précédemment proposées par (Brouwers *et al.*, 2012; Gala & Ziegler, 2016).

3 Simplification de textes

De la méthode manuelle classique aux méthodes automatiques, la simplification a connu une forte croissance, en termes de pertinence des résultats obtenus et de champs d'applications pour lesquels de tels systèmes ont été conçus.

3.1 Simplification manuelle

La simplification manuelle de textes adopte deux grandes approches pour simplifier un texte, la première consiste à créer un texte en suivant des recommandations linguistiques et textuelles destinées à faciliter la compréhension en fonction du type de public visé (la collection « La traversée », par exemple, s'adresse à des adultes débutants en lecture ou faibles lecteurs). La deuxième approche consiste à transformer un texte original, jugé difficile à comprendre, en un texte plus simple, destiné à un public particulier. L'exemple le plus connu est Vikidia, un ensemble de textes encyclopédiques destinés aux enfants et directement inspiré des articles de Wikipédia. Plusieurs études ont souligné

Niveau	Phénomène	Explication
Lexical	Termes issus d'une langue étrangère	Termes anglais dans un texte français
	Termes difficiles	Termes hors vocabulaire
	Termes non pertinents à la compréhension	Adjectifs, adverbes etc.
Syntaxique	Temps de conjugaison moins courants et plus littéraires	Imparfait, plus que parfait, passé simple etc.
	Éléments grammaticaux secondaires ou redondants	Propositions subordonnées, adverbiales, relatives, compléments circonstanciels
	Structures syntaxiques complexes	Forme négative, discours indirect, forme passive (forme non canonique)
Discursif	Organisation compliquée de l'information	Pour des raisons esthétiques ou autres
	Cohérence et cohésion	Anaphores (nominales et pronominales), anaphores difficiles (chaînes anaphoriques, dialogues etc.)
	Informations secondaires	Trop d'exemples, de définitions etc.
	Manque d'informations	Manque d'exemples, de définitions etc.

TABLE 1 – Typologie des phénomènes problématiques en ST

le rôle de la simplification manuelle de textes pour la compréhension tel que(Anderson & Davison, 1986) etc.

3.2 Simplification automatique

La simplification automatique de textes (SAT) comporte deux tâches principales : la simplification lexicale (SL) et la simplification syntaxique (SS). Celles-ci peuvent être traitées séparément ou conjointement. La SL modifie le vocabulaire du texte en substituant par exemple les termes jugés difficiles par des synonymes ou paraphrases qui sont plus simples à comprendre. Par exemple, la phrase « *parce que sa femme le **gouvernait** entièrement* » pourra être remplacée par « *parce que sa femme le **dirigeait** entièrement* ». D'autre part, la SS a pour objectif de convertir les phrases contenant des structures syntaxiques complexes tels que les propositions subordonnées et les phrases en forme passive en phrases plus simples structurellement en préservant leur sens original (ou au moins en limitant son altération). Par exemple, « *Il reconnut que le cinquième compartiment avait été envahi par la mer* » ; pourra être remplacée par « *Il reconnut que la mer a envahi le cinquième compartiment* ». La SAT comporte trois principales taches : (i) la détection automatique des phénomènes problématiques et des éléments linguistiques complexes impliquant des aspects lexicaux, syntaxiques ou autres (cf. Table 1) (ii) la production d'une version simplifiée des textes à l'aide d'un ensemble d'opérations de simplification (iii) l'évaluation des simplifications apportées.

Dans le domaine de SAT, la majorité des travaux se limite aux traitement de phénomènes problématiques lexicaux ou syntaxiques. Très peu d'attention a été portée au niveau discursif.

4 Approches de simplification

Les approches de SAT peuvent être regroupées en trois grandes familles, notamment les approches par règles classiques, les approches statistiques et finalement la combinaison des deux donnant naissance aux approches hybrides.

4.1 Approches par règles

Ce sont les toutes premières approches adoptées en SAT, elles ont été proposées pour des cas d'application spécifiques et pour une population bien ciblée. (Inui *et al.*, 2003) ont réalisé un système de simplification pour les personnes sourdes, en considérant les deux niveaux lexical et syntaxique. (Williams & Reiter, 2008) ont proposé un système de génération de textes simplifiés au niveau discursif appelé SKILL-SUM pour les personnes avec faible niveau d'alphabétisation, en favorisant les mots et les phrases courtes qui sont plus lisibles et plus compréhensibles par les lecteurs .

La simplification s'effectue sur la base de règles définies explicitement par des experts, en analysant des textes originaux et leurs équivalents simplifiés. Par exemple, (Brouwers *et al.*, 2014) ont utilisé 19 règles pour effectuer une simplification syntaxique sur des textes français, classés en 3 catégories : Suppression (12 règles), Modification (3 règles) et Division (4 règles). La suppression par exemple procède à l'élimination directe des éléments syntaxiques secondaires comme les propositions subordonnées. Exemple : la phrase « *Le candidat a proposé une idée intéressante, qui parait très originale.* » devient « *Le candidat a proposé une idée intéressante* ».

Bien qu'elle soit la plus ancienne, cette approche reste d'actualité notamment pour les langues peu dotées pour lesquelles il n'existe pas de corpus parallèles. Elle sont très adaptées à la simplification syntaxique. Leur principal inconvénient est toutefois une portabilité et une évolutivité réduites pour les nouveaux scénarios, qui nécessitent la création de nouveaux ensembles de règles à chaque fois qu'une nouvelle langue (ou un nouveau domaine) doit être couvert. De plus :

1. Elles consomment beaucoup de temps.
2. Elles requièrent beaucoup d'implication humaine pour la définition des règles.
3. Il est impossible de trouver et énumérer toutes les règles de simplification.

4.2 Approches statistiques

Ces approches reposent principalement sur la disponibilité de grands corpus utilisés à la place de connaissances expertes. L'objectif est d'apprendre les règles automatiquement depuis ces ressources de données.

L'approche de simplification par **transduction d'arbres** (Tree Transduction en anglais) vise à sur-générer des règles de simplification automatiquement, ensuite à choisir celles qui correspondent le mieux. (Paetzold & Specia, 2013) proposent une approche composée de trois modules : Un module d'entraînement qui sur-génère des règles de transformation candidates, pour cela il reçoit en entrée les représentations sous forme d'arbres syntaxiques du corpus parallèle aligné au niveau des mots, effectue les différentes transformations possibles sur l'arbre représentant la phrase complexe pour reproduire l'arbre représentant la phrase simplifiée, ce qui nous donne en sortie les différentes règles candidates (lexicales, syntaxiques, lexico-syntaxiques). Le module sélectionne ensuite les

transformations les plus susceptibles de représenter des opérations de simplification en vérifiant certains critères. Le module de simplification génère à partir d'une phrase complexe en entrée et des règles précédemment sélectionnées, des phrase simplifiées candidates. Finalement, le module de classement attribue des scores aux différentes candidates et les ordonne pour sélectionner la meilleure.

Les approches par **traduction automatique (TA)** (**MT** pour Machine Translation) considèrent la simplification de texte comme un problème de traduction monolingue. Initialement, elles ont été proposées pour faire de la traduction d'une langue en une autre, plus tard elles ont été utilisées pour faire la simplification de texte (Specia, 2010). Le début des années 90 a vu le lancement des approches de traduction automatique statistique (SMT pour Statistical machine translation) où les règles ainsi que les modèles de simplification peuvent être appris automatiquement à partir de corpus parallèles constitués de paires de phrases (Complexes-Simplifiées). Elles regroupent les méthodes suivantes :

— **Traduction automatique statistique fondée sur les mots** (Word Based Statistical Machine Translation) est la plus ancienne de ces méthodes et la moins utilisée, l'unité fondamentale étant le mot, les résultats obtenus étaient moins corrects vu qu'elle effectue une traduction mot à mot en ignorant l'aspect syntaxique et l'aspect sémantique bien entendu.
— **Traduction automatique statistique fondée sur les syntagmes** (Phrase Based Statistical Machine Translation) Le but étant de manipuler des séquences de mots (de taille variable) lors de la traduction à la place de mots seuls. La méthode consiste à segmenter une phrases en syntagmes, ensuite à faire la traduction syntagme par syntagme puis les réordonner pour formuler une phrase de sortie a priori simplifiée. Certains travaux avaient pour objectif la fragmentation des phrases longues en phrases plus courtes et plus simples (Specia, 2010), d'autres se sont intéressés à la suppression des expressions secondaires (Coster & Kauchak, 2011).
Cette méthode ne peut effectuer qu'un petit nombre d'opérations de simplification, telles que la substitution lexicale, la suppression et l'explication simple. Elle n'est pas bien adaptée aux opérations de réorganisation ou de fragmentation.
— **Traduction automatique statistique fondée sur la syntaxe** (Syntax Based Statistical Machine Translation) La différence avec la méthode précédente est qu'elle manipule des unités syntaxiques complètes au lieu de mots ou de syntagmes seuls, incorporant ainsi une représentation explicite de la syntaxe dans les systèmes MT (comme l'ordre des mots par exemple), et permettant donc d'effectuer de meilleures opérations de réorganisations. Exemples de travaux : (Zhu *et al.*, 2010; Xu *et al.*, 2016).

L'inconvénient principal des méthodes de traduction automatique statistiques est qu'elles fonctionnent séparément sur de petits composants de simplification (lexical seulement ou syntaxique seulement). De plus, nous trouvons que parfois, elles traitent partiellement un niveau linguistique, comme la fragmentation uniquement par rapport au niveau syntaxique.

La **Traduction automatique neuronale** (NMT pour Neural Machine Translation) est récemment apparue (Kalchbrenner & Blunsom, 2013) en tant qu'une nouvelle approche de TA, et a montré une forte amélioration par rapport aux approches précédentes classiques. De plus, contrairement aux méthodes traditionnelles qui fonctionnent séparément sur de petits composants, NMT réalise une simplification de bout en bout (end-to-end), cela veut dire qu'elle ne nécessite pas des décodeurs

externes, des modèles de langage [1] etc.
L'approche centrale de NMT est l'architecture d'encodeur-décodeur implémentée par les réseaux de neurones récurrents (RNN), la séquence d'entrée est ainsi représentée par un vecteur (encodage), puis décodée pour obtenir la séquence de sortie représentant une phrase simplifiée. Par exemple, (Zhang & Lapata, 2017) ont introduit une architecture encodeur-décodeur à base de RNN, couplée à un modèle d'apprentissage par renforcement. L'encodeur-décodeur est considéré comme Agent du modèle d'apprentissage par renforcement. Etant donné une phrase complexe en entrée, l'encodeur la transforme en une séquence d'états cachés avec un réseau de neurones LSTM (Long Short Term Memory), et à chaque étape t il réalise une action $\hat{y}t$ appartenant à un vocabulaire fixe de sortie, suivant une certaine politique. L'agent continue à faire des actions, jusqu'à ce qu'il produise la fin de la phrase. La sortie simplifiée de l'encodeur-décodeur serait ainsi $\hat{Y} = (\hat{y}1, \hat{y}2, \hat{y}3 \ldots)$. Une récompense r est alors attribuée, et l'algorithme de renforcement met à jour l'agent. Y est la sortie 'Référence', elle est utilisée pour l'évaluation de la simplicité de la sortie simplifiée générée et le calcul de la récompense r.

Un problème avec toutes les approches statistiques, est que la pertinence des résultats dépend largement de la taille des corpus parallèles utilisés, dont leur construction est en général très coûteuse. Certaines solutions ont été proposées pour contourner ce problème de manque de ressources notamment le travail de (Aprosio *et al.*, 2019) où ils ont construit un nouveau modèle encodeur-décodeur basé sur l'attention (Vaswani *et al.*, 2017). Leur idée repose sur l'extension de petits corpus d'entraînement avec des données synthétiques pour satisfaire le besoin en données volumineuses nécessaires à l'entraînement de leur modèle. Le modèle de simplification fournit une solution de bout en bout pour traiter à la fois la simplification lexicale et syntaxique, ceci en apprenant à faire des changements structurels plus complexes que les changements terme à terme uniquement.

(Aprosio *et al.*, 2019) ont suivi trois stratégies différentes pour l'augmentation (extension) des données, à partir des données « de référence » (gold) à leur disposition et qui sont en quantité limitée. La première stratégie est le sur-échantillonnage, celle-là consiste à faire augmenter la taille du corpus d'entraînement en dupliquant le corpus original plusieurs fois afin de maximiser l'exploitation des paires de phrases références à leur disposition. La deuxième consiste à créer des paires synthétiques simple-simple à partir de grands corpus monolingues, ceci est réalisé en extrayant automatiquement les phrases les plus simples avec des méthodes heuristiques, ensuite les dupliquer pour créer des paires simple-simple, celles-ci sont ensuite utilisées comme données synthétiques pour former le système de simplification. L'objectif étant d'introduire un biais dans le décodeur en vue d'améliorer la simplicité des sorties. La troisième stratégie est la création de paires synthétiques complexe-simple pour former un système complexifiant qui génère des phrases complexes à partie de celles simples précédemment sélectionnées (en utilisant des outils d'Open-NMT). Les paires sont ensuite inversées pour former le système de simplification. L'intuition étant de conserver les phrases simplifiées générées par l'expertise humaine dans la partie cible des données parallèles afin d'améliorer la génération de phrases simplifiées.

Leur système de simplification fonctionne de la manière suivante : Initialement, une séquence de mots est transmise à l'encodeur. A chaque pas de temps, sur la base des représentations générées par l'encodeur et du mot généré dans le pas de temps précédent, le décodeur génère le mot suivant. Ce

1. Un modèle de langage correspond à une fonction, qui prend une phrase traduite et renvoie la probabilité qu'elle soit dite par un locuteur natif, de plus elle peut aider à choisir une traduction parmi plusieurs pour un mot donné, selon son contexte.

processus se poursuit jusqu'à ce que le décodeur génère le symbole de fin de phrase. Un module de génération de pointeurs est rajouté à ce modèle, il permet à la fois de copier des mots à partir de la phrase source et de générer des mots à partir d'un vocabulaire fixe partagé contenant tous les mots des phrases d'apprentissage complexes et simples. À chaque pas de temps, le réseau estime la probabilité de générer un mot et utilise cette probabilité pour décider de générer ou de copier le mot.

Récemment, des techniques impliquant de l'apprentissage non supervisé ont vu le jour. (Surya *et al.*, 2018) ont conçu un modèle avec un encodeur partagé et deux décodeurs à base d'attention, en plus d'un discriminateur pour influencer le comportement du décodeur et d'un classifieur pour la diversification, selon la nature de l'entrée (simple/complexe). Le modèle reconstitue d'un coté l'entrée originale complexe à partir de la phrase simplifiée, et de l'autre génère une version simplifiée de l'entrée. Ceci est réalisé en examinant la structure et les schémas linguistiques d'un grand nombre de phrases simples et complexes non alignées (à partir de Wikipédia) qui sont beaucoup moins coûteuses et faciles à obtenir que les données parallèles alignées.

4.3 Approches hybrides

L'objectif de ces approches est de tirer profit des avantages des deux approches précédentes (par règles et statistiques) en combinant la simplification syntaxique par règles, et la simplification lexicale par approches statistiques. Ceci en fait est dû aux limitations de la simplification lexicale basée sur les règles (trop de règles à énumérer manuellement), et celles de la simplification syntaxique statistique (qui produit des résultats moins bons grammaticalement et qui sont très limitées dans leur portée, comme le cas du passage de la voix passive à la voix active, chose qui est beaucoup plus pertinente par règles). Un exemple de travaux est celui de (Siddharthan & Mandya, 2014), qui ont proposé un système hybride combinant un module de simplification lexicale statistique, et un module de simplification syntaxique basée sur 136 règles.

Les approches neuronales sont considérés plus performantes, en effet elles traitent la simplification de textes de bout en bout, d'une façon plus rapide et en donnant des résultats plus pertinents contrairement aux méthodes hybrides modulaires où les résultats sont relativement moins pertinents pour certaines opérations.

5 Evaluation des systèmes de simplification de textes

L'évaluation des systèmes de SAT vise généralement à tester leur pertinence en termes de qualité des sorties, ou à évaluer leur efficacité/utilité en mesurant le temps de lecture et la compréhension des lecteurs (utilisateurs finaux). Cependant, ce n'est pas toujours possible de procéder à un test réel effectué par la population cible comme les dyslexiques où les apprenants de langue. Une évaluation par des experts du domaine est alors effectuée, en attribuant des scores sur la simplicité des résultats, leur pertinence en termes grammaticalité, et leur préservation du sens. Cette méthode est dite manuelle ou humaine, et elle n'opère qu'au niveau de la phrase. Des méthodes automatiques sont alors apparues pour faciliter la tâche aux concepteurs de systèmes de simplification, en leur économisant du temps et de coût. Ces évaluations automatiques s'appliquent sur le texte entier, en s'appuyant par exemple sur des mesures dédiées aux systèmes MT ou des formules spécifiques. La figure 1 montre un schéma synthétisant les techniques l'évaluation existantes.
Il est à noter que l'évaluation s'effectue sur les textes simplifiés aussi bien que sur les textes originaux

complexes. Pour le premier cas comme nous l'avons mentionné plus haut, elle mesure la lisibilité des textes pour tester l'efficacité du système de simplification et la qualité de ces sorties. La deuxième vise à identifier à partir du texte original, les parties qui sont particulièrement complexes et qui doivent par conséquent être simplifiées.
La qualité du résultat généré par les systèmes de simplification de textes est généralement évaluée en utilisant une combinaison de mesures de lisibilité automatiques (mesure du degré de simplicité principalement) et d'évaluation humaine (mesure de grammaticalité, de simplicité et de préservation du sens).

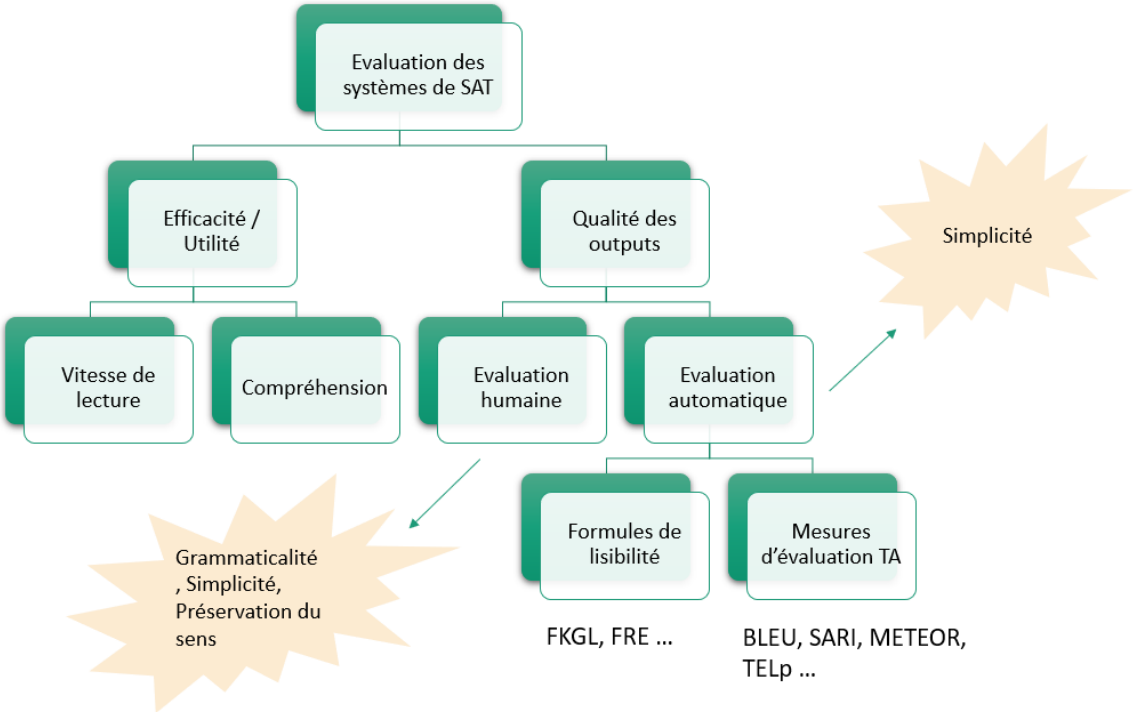

FIGURE 1 – Évaluation des systèmes de SAT. Schéma inspiré de (Štajner & Saggion, 2018)

5.1 Évaluation manuelle

L'évaluation manuelle (EM) se fait en mesurant certains critères sur les textes simplifiés en sortie à savoir : l'aisance (ou la facilité) de lecture qui mesure l'exactitude et la grammaticalité, le degré de simplicité et la préservation du sens qui mesure la correspondance du sens de la phrase simplifiée à celui de la phrase originale. En général, ces mesures sont données sur une échelle de 1 à 5. Plus le score est élevé, plus la qualité du texte simplifié en sortie est meilleure.

5.2 Évaluation automatique

Comme pour les nouvelles méthodes de SAT, leur évaluation en parallèle a eu sa part d'automatisation. En effet, l'EM présente quelques inconvénients, notamment le coût élevé, la consommation du temps, le manque d'adaptabilité, et la non-reproductibilité. L'évaluation automatique (EA) vient alors non pas pour remplacer l'EM mais plutôt pour la compléter. En effet, parfois elle ne mesure que le degré

de simplicité des sorties, contrairement à l'EM qui mesure la grammaticalité, la préservation du sens et la simplicité. Elle s'applique de l'une des manières suivantes :

Formules de lisibilité Depuis 1950, plus de 200 formules d'évaluation ont été développées pour l'anglais, or ce n'était pas le cas pour le Français où on ne retrouve que quelques travaux comme celui de (François & Watrin, 2011). Les toutes premières formules ont été calculées sur la base de la longueur des mots et la longueur moyenne des phrases seulement, parce que ces deux paramètres se corrèlent bien avec les tests de lecture, parmi ces formules : FKGL 'Flesch-Kincaid Grade Level index' (Kincaid *et al.*, 1975) ; FRE 'Flesch Reading Ease score' (Flesch & Gould, 1949) ; Ces formules sont restées largement utilisées jusqu'à maintenant. D'autres formules comme Dale-Chall (Dale & Chall, 1948) ont été calculées sur la base de la longueur moyenne des phrases et de la proportion de mots hors le vocabulaire simple.
Cependant, ces formules présentent plusieurs lacunes. Par exemple, elles ne prennent en compte que les caractéristiques superficielles, en ignorant d'autres aspects importants contribuant à la difficulté du texte, tels que la cohérence, la densité du contenu, la capacité d'inférence [2] du lecteur etc. Ils omettent également l'aspect interactif du processus de lecture.

L'avancement puissant en méthodes du TAL et des techniques du ML ont fait émerger de nouvelles approches pour l'évaluation de la lisibilité qui donnent de meilleurs résultats par rapport aux formules précédentes. (Petersen & Ostendorf, 2009) ont utilisé des modèles de langues [3] et des SVM avec de nouveaux attributs tels que la hauteur moyenne de l'arbre syntaxique, le nombre moyen de phrases nominales et verbales, nombre moyen de syllabes par mot etc. (Feng *et al.*, 2010), a montré que d'autres attributs tels que la longueur des chaînes lexicales, la densité des entités nommées dans le document, la longueur moyenne des phrases (qui est plus utile et moins coûteuse), se corrèlent mieux que le FKGL avec la compréhension du lecteur, notamment les apprenants d'une langue étrangère et les enfants avec troubles de lecture. (François & Fairon, 2012) ont présenté une nouvelle formule d'évaluation de la lisibilité pour 'le Français comme langue étrangère (FLE)', en utilisant les SVMs et 46 caractéristiques textuelles couvrant les trois niveaux lexical, syntaxique et sémantique, ainsi que des caractéristiques spécifiques au FLE tels que la fréquence moyenne des expressions multi-mots dans le texte, la nature du texte qui s'obtient en calculant certains indicateurs (le taux de ponctuation, présence des virgules etc.). (Vajjala & Meurers, 2014) ont utilisé un modèle de régression entraîné sur des documents entiers pour comparer la lisibilité des phrases parallèles alignés du corpus Wikipédia-SimpleWikipédia (Zhu *et al.*, 2010).

Mesures des systèmes de traduction automatique En raison du développement large des systèmes de simplification de TA, il est devenu de plus en plus indispensable d'évaluer leurs performances, non seulement pour établir une comparaison, mais aussi pour savoir s'ils réalisaient des progrès. Cependant, l'évaluation de la traduction automatique est difficile car les langues naturelles sont très ambiguës et le contenu s'exprime différemment d'une langue à une autre.
Les techniques actuelles se basent principalement sur la comparaison entre les résultats produits par les systèmes de simplification et les textes simplifiés manuellement considérés 'de référence', ici on distingue différentes manières de faire : soit en fournissant une seule référence ou bien en donnant des références multiples pour améliorer la précision de la comparaison. D'autres mesures n'utilisent

2. Une opération mentale qui permet au lecteur de déduire les non-dits ou les éléments implicites dans un texte
3. Prédisent la probabilité qu'une séquence de mots particulière se produise

pas de textes de référence. En général, les mesures des systèmes MT évaluent la **similitude lexicale** par le moyen de l'ordre des mots dans les phrases, la distance de modification, et le chevauchement des séquences des mots. Des **caractéristiques linguistique** sont également pris en compte, telles que la syntaxe et la sémantique (l'étiquetage morpho-syntaxique ou POS-tagging, les structures des phrases, les synonymes, les entités nommées, les rôles sémantiques et les modèles de langage etc.) Ces deux catégories sont généralement inséparables, en effet certaines mesures de la première catégorie utilisent certaines caractéristiques linguistiques et vice versa. Les dernières recherches appliquent des modèles d'**apprentissage profond** pour l'évaluation . Nous présentons en particulier certaines mesures d'évaluation pour la similarité lexicale :

1. **Distance des modifications**
   Elle se calcule en comptant le nombre minimum de changements à apporter au texte simplifié pour le transformer en texte de référence. On trouve :
   — WER (Word Error Rate) : introduit par (Su *et al.*, 1992), elle tient compte de l'ordre des mots. Les opérations possibles sont l'insertion, la suppression et la substitution des mots. Le nombre minimum des modifications est calculé par la formule suivante :

$$WER = \frac{N(Substitutions) + N(Insertions) + N(Suppressions)}{LongueurRef}$$

   tel que : N(x) est le nombre d'opérations x effectuées et LongueurRef est la longueur de la phrase de référence. Une des faiblesses de WER est qu'elle ne tient pas compte de l'ordre des mots correctement, en effet elle est très réduite lorsque l'ordre d'un mot ne correspond pas à celui du texte de référence (Fausses phrases).
   — PER (Position-independent Word Error Rate) : proposée par (Tillmann *et al.*, 1997) pour résoudre ce problème de fausses phrases, en ignorant l'ordre des mots lors de la reconstitution de la phrase de référence. La mesure se fait en calculant le nombre de fois qu'un mot apparaît identiquement dans les deux phrases (simplifié-référence) qui est donné par le paramètre 'Correct', et selon la différence en longueur entre les deux phrases, le reste des mots sont insérés ou supprimés. Voici la formule correspondante :

$$PER = 1 - \frac{Correct - Max(0\ ,\ LongSortie - LongRef)}{LongRef}$$

2. **Précision et rappel**
   Ce sont deux propriétés qui ont été confirmées par plusieurs mesures comme étant essentielles pour une corrélation élevée avec les jugements humains. Nous introduisons :
   — BLEU (Bilingual Evaluation Understudy) : est la plus utilisée et la moins coûteuse, proposée par (Papineni *et al.*, 2002). Elle montre une corrélation élevée avec les jugements humains pour la grammaticalité et la préservation du sens, elle n'est cependant pas bien adaptée pour l'évaluation de la simplicité des résultats. Le calcul du score se fait en premier lieu au niveau de chaque segment du texte simplifié (généralement une phrase), ensuite un score total correspondant à la moyenne géométrique est calculé, elle est basée sur le calcul de la précision pour des n-grammes de taille 1 à 4 avec un coefficient de pénalité de brièveté (BP), et donne des poids égaux pour les différents n-grammes. Le résultat est toujours compris entre 0 et 100 %, plus il est proche de 100, meilleure est la simplification. Cette mesure pénalise fortement la réorganisation des mots et le raccourcissement des phrases. Sa formule étant :

$$BLEU = BP * \exp \sum_{i=1}^{n} \lambda_n \log Precision_n$$

$$BP = \begin{cases} 1 & \text{si } c > r \\ \exp^{(1-r/c)} & \text{si } c \leq r \end{cases}$$

Où **"c"** est la longueur totale de la phrase simplifiée ; **"r"** la longueur de la phrase de référence, et si plusieurs références existent, celle de longueur la plus proche à celle de la phrase simplifiée est choisie ; λ_n sont des poids positifs de précision pour les n-grammes (leur somme est à 1) généralement pris identiques.

— SARI : proposée récemment par (Xu *et al.*, 2016). Elle mesure la simplicité à travers des opérations d'ajout, de suppression et de conservation. Elle se différencie par le fait que la phrase simplifiée est comparée aux phrases de référence ainsi que la phrase d'origine non simplifiée. SARI a montré une forte corrélation avec les jugements humain pour la simplicité et est actuellement la principale mesure utilisée pour évaluer les modèles de simplification. Sa formule :

$$SARI = c1 * F_{ajout} \ + \ c2 * F_{conservation} \ + \ c3 * F_{suppression}$$

$$F_{oper} = \frac{2*P_{oper}*R_{oper}}{R_{oper} + P_{oper}} \ \text{|} \ P_{oper} = \frac{1}{k} * \sum_{n=1}^{k} p_{oper(n)} \ \text{|} \ R_{oper} = \frac{1}{k} * \sum_{n=1}^{k} r_{oper(n)}$$

Tel que : c1 = c2 = c3 = 1/3 ; **"k"** représente l'ordre le plus grand des n-grammes ; $oper \in \{ajout, conservation, suppression\}$; $p_{oper(n)}$ et $r_{oper(n)}$ représentent respectivement la précision et le rappel des n-grammes correspondants à l'opération en question.

— F-mesure : c'est une combinaison de la précision P et du rappel R, elle était d'abord adoptée par la recherche d'information, ensuite par l'extraction d'information et l'évaluation des systèmes de simplification TA. La formule utilisée étant :

$$F_\beta = (1 \ + \ \beta^2) \ \frac{P*R}{R + \beta^2 * P}$$

3. **Ordre des mots**
L'ordre des mots est un facteur très significatif pour évaluer la similarité lexicale. La diversité linguistique permet toutefois différentes apparences ou structures pour une phrase, le défi est donc de pouvoir appliquer la pénalité sur les mots qui sont vraiment incorrects (des phrases mal structurées) plutôt que sur des mots corrects avec un ordre différent. Au contraire, une phrase simplifiée ayant un ordre de mots différent de celui de la référence mais correcte est un vrai point d'intérêt, vu que ça permettra éventuellement une meilleure représentation de l'information dans la phrase d'une manière pouvant être plus claire est plus simple que dans la phrase référence. On trouve les mesures : ATEC (Assessment of Text Essential Characteristic), PORT (Precision-Order-Recall MT Evaluation Metric for Tuning), LEPOR (Length Penalty, Precision, n-gram Position difference Penalty and Recall) etc.

6 Conclusion et perspectives

Nous avons présenté à travers cet article, l'état de l'art en simplification de texte. Cette dernière est définie comme un processus permettant de modifier des textes difficiles afin de les rendre plus simples et plus accessibles à certaines catégories de lecteurs. Nous avons élaboré en premier lieu une typologie de phénomènes problématiques comprenant trois niveaux linguistiques (Lexical, syntaxique

et discursif). Ensuite nous avons abordé les différents types d'approches pour la simplification, notamment l'approche par règles, l'approche statistique et l'approche hybride. Les approches statistiques consistent actuellement à faire en grande partie de l'apprentissage automatique/profond qui prennent de plus en plus de l'ampleur en TAL, notamment en simplification de textes. Finalement, nous avons introduit certaines mesures d'évaluation pour les systèmes de simplification telles que les formules de lisibilité et les mesures des systèmes MT.
La simplification de textes implique toutefois des problèmes majeurs liés à l'apprentissage et à la compréhension, vu qu'elle introduit des erreurs et des ambiguïtés lors de la modification du texte original. De plus, une simplification excessive risque de limiter l'apprentissage et la transmission d'informations. Nous travaillons par conséquent sur un nouveau paradigme que nous appelons l'explicitation de textes, et qui permet de ne pas modifier le texte original. Cela consiste en revanche à adapter les approches de simplification de bout en bout (qui intègrent à la fois la détection et la simplification des difficultés), de sorte à faire la détection mais à traiter les phénomènes problématiques différemment, en enrichissant le texte original d'éléments explicitant ses informations implicites et difficiles. Les architectures récentes à base de transformeurs peuvent être très intéressantes dans le cadre de l'explicitation, et notre travail consiste à proposer un tel système d'explicitation pour deux catégories de lecteurs (Les enfants dyslexiques et les apprenants du français comme langue étrangère).

Remerciements

Nous remercions Solen QUINIOU et Béatrice DAILLE pour leur encadrement et leur appui scientifique. Nous remercions Lynda SAID LHADJ pour son support et pour le partage de savoir-faire et de connaissances. Merci à tous les trois pour la relecture de l'article.

Références

ALUÍSIO S. M. & GASPERIN C. (2010). Fostering digital inclusion and accessibility : the porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, p. 46–53 : Association for Computational Linguistics.

ANDERSON R. C. & DAVISON A. (1986). Conceptual and empirical bases of readability formulas. *Center for the Study of Reading Technical Report ; no. 392*.

APROSIO A. P., TONELLI S., TURCHI M., NEGRI M. & DI GANGI M. A. (2019). Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, p. 37–44.

BROUWERS L., BERNHARD D., LIGOZAT A.-L. & FRANÇOIS T. (2012). Simplification syntaxique de phrases pour le français (syntactic simplification for french sentences)[in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 211–224.

BROUWERS L., BERNHARD D., LIGOZAT A.-L. & FRANÇOIS T. (2014). Syntactic sentence simplification for french. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, p. 47–56.

CARROLL J., MINNEN G., CANNING Y., DEVLIN S. & TAIT J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, p. 7–10.

CHANDRASEKAR R., DORAN C. & SRINIVAS B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, p. 1041–1044 : Association for Computational Linguistics.

COSTER W. & KAUCHAK D. (2011). Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation*, p. 1–9 : Association for Computational Linguistics.

DALE E. & CHALL J. S. (1948). A formula for predicting readability : Instructions. *Educational research bulletin*, p. 37–54.

FENG L., JANSCHE M., HUENERFAUTH M. & ELHADAD N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd international conference on computational linguistics : Posters*, p. 276–284 : Association for computational linguistics.

FLESCH R. & GOULD A. J. (1949). *The art of readable writing*, volume 8. Harper New York.

FRANÇOIS T. & FAIRON C. (2012). An ai readability formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 466–477 : Association for Computational Linguistics.

FRANÇOIS T. & WATRIN P. (2011). Quel apport des unités polylexicales dans une formule de lisibilité pour le français langue étrangère. *Traitement Automatique des Langues Naturelles*, p.49̃.

GALA N., FRANÇOIS T., JAVOUREY-DREVET L. & ZIEGLER J. C. (2018). La simplification de textes, une aide à l'apprentissage de la lecture. *Langue française*, (3), 123–131.

GALA N. & ZIEGLER J. (2016). Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia. In *Computational Linguistics for Linguistic Complexity, workshop at COLING (Computational Linguistics conference)*, Osaka, Japan. HAL : hal-01757941.

INUI K., FUJITA A., TAKAHASHI T., IIDA R. & IWAKURA T. (2003). Text simplification for reading assistance : a project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, p. 9–16 : Association for Computational Linguistics.

KALCHBRENNER N. & BLUNSOM P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1700–1709.

KINCAID J. P., FISHBURNE JR R. P., ROGERS R. L. & CHISSOM B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

PAETZOLD G. H. & SPECIA L. (2013). Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.

PETERSEN S. E. & OSTENDORF M. (2009). A machine learning approach to reading level assessment. *Computer speech & language*, **23**(1), 89–106.

QUINIOU S. & DAILLE B. (2018). Towards a Diagnosis of Textual Difficulties for Children with Dyslexia. In *11th International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan. HAL : hal-01737726.

SAGGION H., ŠTAJNER S., BOTT S., MILLE S., RELLO L. & DRNDAREVIC B. (2015). Making it simplext : Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, **6**(4), 1–36.

SERETAN V. (2012). Acquisition of syntactic simplification rules for french.

SIDDHARTHAN A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, **4**(1), 77–109.

SIDDHARTHAN A. & KATSOS N. (2010). Reformulating discourse connectives for non-expert readers. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 1002–1010 : Association for Computational Linguistics.

SIDDHARTHAN A. & MANDYA A. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, p. 722–731.

SPECIA L. (2010). Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language*, p. 30–39 : Springer.

ŠTAJNER S., EVANS R., ORASAN C. & MITKOV R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of the the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, p. 14–21 : Citeseer.

ŠTAJNER S. & SAGGION H. (2018). Data-driven text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics : Tutorial Abstracts*, p. 19–23.

SU K.-Y., WU M.-W. & CHANG J.-S. (1992). A new quantitative quality measure for machine translation systems. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, p. 433–439 : Association for Computational Linguistics.

SURYA S., MISHRA A., LAHA A., JAIN P. & SANKARANARAYANAN K. (2018). Unsupervised neural text simplification. *arXiv preprint arXiv :1810.07931*.

TILLMANN C., VOGEL S., NEY H., ZUBIAGA A. & SAWAF H. (1997). Accelerated dp based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology*.

VAJJALA S. & MEURERS D. (2014). Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, p. 288–297.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in neural information processing systems*, p. 5998–6008.

WILLIAMS S. & REITER E. (2008). Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, **14**(4), 495–525.

XU W., NAPOLES C., PAVLICK E., CHEN Q. & CALLISON-BURCH C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, **4**, 401–415.

ZHANG X. & LAPATA M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 584–594, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1062.

ZHU Z., BERNHARD D. & GUREVYCH I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, p. 1353–1361 : Association for Computational Linguistics.