Simplification interactive pour la production de textes adaptés aux personnes souffrant de troubles de la compréhension

Aurélien Max
LIR-LIMSI-CNRS
Université Paris 11
Orsay, France
aurelien.max@limsi.fr

Mots-clefs : troubles du langage, simplification syntaxique, règles de réécriture, validation interactive, traitements de texte

Keywords: language disorders, syntactic simplification, rewriting rules, interactive validation, word processors

Résumé Cet article traite du problème de la compréhensibilité des textes et en particulier du besoin de simplifier la complexité syntaxique des phrases pour des lecteurs souffrant de troubles de la compréhension. Nous présentons une approche à base de règles de simplification développées manuellement et son intégration dans un traitement de texte. Cette intégration permet la validation interactive de simplifications candidates produites par le système, et lie la tâche de création de texte simplifié à celle de rédaction.

Abstract This paper addresses the issue of text readability and in particular the need for simplifying the syntactic complexity of sentences for language-impaired readers. The proposed approach uses handcrafted simplification rules and has been integrated into a word processor. This allows interactive validation of candidate simplified sentences produced by the system, and integrates the task of creating simplified texts into that of authoring.

1 Simplification de texte

La *compréhensibilité* est une propriété cruciale d'un texte, et l'usage d'une langue sans contraintes particulières peut mener à des textes difficiles à comprendre. Or la vocation première d'un texte informatif est de véhiculer une information auprès de ses lecteurs (soit d'atteindre certains *buts communicatifs*), ce qui impose des contraintes sur la rédaction de ce texte. Des troubles du langage peuvent néanmoins interdire la compréhension de textes qui seraient autrement jugés raisonnablement compréhensibles. L'*aphasie de Broca*[1] a pour effets de rendre compliquée l'expression ainsi que l'interprétation d'énoncés contenant une certaine *complexité*

[1]L'aphasie est un trouble du langage pouvant survenir à la suite d'une attaque cérébrale touchant les lobes frontaux et temporaux de l'hémisphère gauche du cerveau. Plusieurs catégories d'aphasies existent, et elles peuvent affecter la compréhension aussi bien que la production du langage.

linguistique, alors que les connaissances sémantiques du sujet sont intactes. Il semble donc important de pouvoir produire des textes adaptés à ces lecteurs, ou au moins de dériver des textes simplifiés à partir de textes existants.

La simplification de texte a récemment été l'objet de plusieurs travaux, avec comme motivation la simplification de texte en vue d'une analyse syntaxique (Chandrasekar et al, 1996) ou la simplification de texte destinée aux lecteurs souffrant de troubles du langage (Carroll et al, 1998; Devlin, 1999; Liben-Nowell, 2000; Canning, 2002). L'approche de Chandrasekar et al. se base sur des règles de simplification syntaxique qui sont apprises depuis un corpus simplifié manuellement. Les autres approches utilisent des règles écrites manuellement qui sont appliquées sur la sortie d'un analyseur syntaxique. Plus récemment, (Siddharthan, 2003) s'est intéressé à la conservation de la cohérence des textes résultant de simplifications syntaxiques.

L'approche que nous présentons ici se distingue des précédentes par le fait qu'elle vise à intégrer un module de simplification de texte à base de règles dans un outil de rédaction traditionnel (un traitement de texte). Il nous semble en effet important de pouvoir considérer la production d'un texte simplifié comme une activité liée à celle de la rédaction du texte dont il est issu, ce qui permet d'impliquer l'auteur dans les choix pour lesquels un système automatique ne pourrait prendre de décision satisfaisante dans le cas général. Dans cet article, nous nous concentrerons sur la simplification syntaxique des phrases[2]. Dans l'approche que nous avons choisie, les simplifications sont le résultat de l'application non-déterministe de règles de réécriture développées manuellement par un linguiste. Cette approche est en partie justifiée par la difficulté à obtenir des corpus alignés de textes et de leur simplification qui pourraient servir à un apprentissage de ces règles[3] ou à des techniques basées sur l'exemple.

2 Système de simplification syntaxique par réécriture

Nous avons choisi une approche à base de règles pour la simplification syntaxique des phrases d'un texte en anglais s'inspirant de travaux sur la transduction d'arbres. L'utilisation de règles permet d'exprimer des conditions sur leur applicabilité qui offrent une meilleure maîtrise du contexte de simplification qui peut être plus ou moins spécifié (des mots jusqu'aux catégories syntaxiques profondes les dominant par exemple). Le développement des règles peut être incrémental, ce qui autorise une évaluation progressive du système sur un corpus de test constant, afin de contrôler que l'ajout de règles ne dégrade pas les performances. Par ailleurs, la non-couverture d'une structure syntaxique particulière ne réalise pas de simplification non souhaitée, et laisse donc localement la phrase inchangée, ce qui garantit au pire la conservation de sa complexité syntaxique. Une telle approche nous a semblé d'autant plus acceptable si elle est intégrée à un traitement de texte qui autorise facilement les révisions ou modifications.

Le niveau de description des phrases utilisé pour décrire les patrons de réécriture est choisi par le linguiste en charge de l'écriture des règles. Ce niveau doit être dérivable automatiquement à partir du texte, ce qui pose la question du compromis entre la robustesse de l'analyse et la finesse de description. Au minimum, la séquence de parties du discours des mots du texte peut

[2]Un système complet de simplification de texte doit également au minimum pouvoir permettre de résoudre les anaphores ainsi que de simplifier la complexité lexicale.

[3]Comme le note (Siddharthan, 2003), la simplification manuelle d'un corpus de texte permettrait vraisemblablement de proposer un jeu de règles de simplification qui devrait avoir une performance au moins comparable à celle de règles apprises sur ce corpus simplifié.

être utilisée comme une structure plate, mais elle n'offre que peu de possibilités de simplification. Des structures syntaxiques plus ou moins profondes seront donc utilisées en fonction des analyseurs disponibles[4].

Une contrainte importante de notre approche de réécriture par règles est que les informations nécessaires à la génération des phrases simplifiées doivent pouvoir être dérivées à partir des phrases en entrée (bien que l'auteur ait ensuite l'opportunité de modifier les phrases simplifiées). Si passer par une représentation sémantique intermédiaire des phrases pour faire de la simplification par regénération serait souhaitable, cela pose d'importants problèmes d'analyse et de représentation et implique la disponibilité d'un moteur de génération de texte paramétrable. Les règles syntaxiques permettent de réutiliser le contenu présent en entrée dans la sortie du simplificateur, et de nouveaux éléments peuvent soit être directement spécifiés dans la sortie, soit exprimés par le biais de conditions sur les règles.

Il n'est cependant pas entièrement possible de s'affranchir de certaines fonctionnalités de génération morphologique, puisque certaines modifications de la syntaxe impliqueront la production de formes de surface adaptées au nouveau contexte[5]. Divers modules d'analyse et de génération linguistiques peuvent ainsi être utilisés dans les conditions des règles.

La simplification d'un texte est opérée phrase par phrase. Les patrons spécifiant les structures syntaxiques à réécrire sont recherchés dans la structure obtenue pour la phrase en entrée, et ce à quelque profondeur que ce soit. Sous réserve que les éventuelles conditions d'application soient remplies, le patron spécifiant la sortie remplace le patron d'entrée, ce qui peut générer autant de simplifications qu'il y a d'occurrences du patron d'entrée dans la phrase. Les patrons d'entrée des règles sont récursivement recherchés dans les sorties du système pour produire l'ensemble des simplifications possibles tenant compte des différents ordres d'application des règles.

Le format des patrons de sortie doit donc être compatible avec le format des patrons d'entrée, mais les étiquettes de sortie peuvent être volontairement modifiées afin, soit d'interdire l'application ultérieure d'autres règles sur un constituant particulier, soit au contraire de déclencher l'application d'une règle réalisant un traitement particulier[6].

Enfin, la nature des réécritures effectuées requiert la possibilité d'insérer dans la sortie du système de nouvelles phrases, qui pourront précéder ou suivre la phrase dans laquelle le patron d'entrée a été trouvé.

Format des règles de réécriture Nous avons donné une importance particulière au fait que les règles puissent être écrites par des linguistes familiers notamment avec des notations d'arbres de dérivation syntaxiques. Les règles sont fortement basées sur la notion d'unification de varia-

[4]Un analyseur probabiliste robuste pour l'anglais tel que RASP (Bricoe et Carroll, 2002) offre une solution robuste pouvant retourner une forêt de solutions annotées par leur probabilité. Cela pose notamment le problème de la sélection de l'analyse retenue pour opérer la simplification. Bien qu'il soit possible de considérer par défaut l'analyse la plus probable retournée par le système, une certaine désambiguïsation parmi les plus probables pourrait être envisagée dans notre contexte interactif.

[5]C'est par exemple le cas du verbe de la proposition principale lors du passage d'une phrase de la voix passive à la voix active, puisque celui-ci doit s'accorder en personne et en nombre avec l'agent du verbe qui était précédemment situé dans un syntagme prépositionnel, ex: *The cat is chased by the dog.* → *The dog* ***chases*** *the cat.*

[6]Par exemple, il est ainsi possible de marquer un groupe nominal indéfini comme devant être transformé en groupe nominal défini par une règle appropriée lorsqu'il est repris dans une nouvelle phrase, ex.:*A terrifying dog chases the cat.* → *A dog is chasing the cat.* ***The dog*** *is terrifying.*

bles, qui peuvent correspondre à des littéraux, à des arbres, ou à des forêts d'arbres[7].

La règle ci-dessous donne un exemple pour le passage à la voix active de phrases à la voix passive contenant un agent exprimé[8]. Les catégories et structures syntaxiques de l'exemple sont celles utilisées dans le Penn TreeBank qui nous a servi comme corpus de test initial.

```
define Activise passive sentences with overt agent

if      [be, InflBe] is analyzeVerb(TagBe, [Be]);
        ?OptAdvs contains only advp rb; ?OptPart contains only prt;
rewrite [s      [       ?Opt1
                        [np-Index NPTheme]
                        ?Opt2
                        [vp     [       [TagBe [Be]]
                                        [vp  [  ?OptAdvs
                                                [vbn Verb]
                                                ?OptPart
                                                [np [[none [Trace-Index]]]]
                                        ?Opt3
                                        [pp     [       [prep [by]]
                                                        [np/lgs NPAgent]]]
                                                        ?Opt4]]]]
                        ?Opt5]]
as      [s      [       ?Opt1
                        [np NPAgent]
                        ?Opt2
                        [vp     [       [SurfaceTag [SurfaceVerb]]
                                        ?OptPart
                                        [np NPTheme]
                                        ?Opt3
                                        ?Opt4]]
                        ?Opt5]]
where   [Number, Person] is number(NPAgent);
        [BaseForm, Infl] is analyzeVerb(vbn, Verb);
        [SurfaceVerb, SurfaceTag] is generateVerb(BaseForm, InflBe, Number, Person);
```

La première condition de la clause **if** qui fait l'appel de la fonction linguistique `analyzeVerb` indique que le verbe de la clause principale doit avoir pour lemme *be*, et que la flexion du verbe (temps et personne) doit se retrouver dans la variable `InflBe`. Cette dernière information sera utilisée dans la clause **where** pour produire la version de surface du verbe principal dans la phrase à la voix active (fonction `generateVerb`). Les deux autres conditions de la clause **where** tentent de reconnaître la personne et le nombre du groupe nominal agent, ainsi que le lemme du verbe au participe passé (fonctions `number` et `analyzeVerb`).

Les constituants optionnels représentant des forêts d'arbres (ex: `?OptAdvs`) peuvent être réutilisés ou non dans les patrons de sortie, mais il n'est pas possible de faire référence à leur structure interne dans la règle dans laquelle ils sont reconnus, ce qui pourra être fait par l'application de règles ultérieures si ces constituants sont réinjectés dans un patron de sortie. Cette fonctionnalité est particulièrement utile pour des constituants qui ne sont pas pertinents pour la règle de simplification en cours d'application mais qui sont dominés par le constituant racine d'un pa-

[7] L'implémentation du moteur de simplification a été réalisée en langage Prolog.

[8] Une estimation (Canning, 2002) indique qu'environ 80% des phrases en anglais au passif n'ont pas d'agent exprimé, et que dans ce cas il est particulièrement difficile, voire impossible, de le retrouver (ex: *She was taken to the hospital*). Notre système n'a pas la prétention de simplifier de telles phrases, bien qu'il serait possible dans certains cas d'insérer un agent indéfini tel que *something* ou *someone* (ce qui requiert néanmoins la détermination du caractère "animé" de cet agent).

tron d'entrée (par exemple, `?Opt4` et `?Opt5` regrouperont l'ensemble des constituants suivant le syntagme prépositionnel contenant l'agent et appartenant au groupe verbal englobant).

Sélection de la meilleure simplification L'application des règles de simplification syntaxique produit une liste de sorties qui sont ordonnées par ordre de production par le moteur de simplification, qui dépend de l'ordre d'analyse des règles. Puisque nous avons choisi d'intégrer la simplification syntaxique dans un traitement de texte, il est possible de demander à l'utilisateur de choisir la simplification qui lui semble la plus appropriée. Cependant, il est souhaitable de pouvoir au préalable ordonner les simplifications candidates par un score décroissant qui indique leur *qualité*. Idéalement, cette qualité impliquerait une prise en compte de la compréhensibilité des phrases, ainsi que de la conservation du sens par rapport à la phrase de départ. Or ce dernier élément relève du jugement expert de l'auteur du texte (ou de l'utilisateur de notre système)[9].

De nombreuses mesures de compréhensibilité (*readability*) des textes ont été proposées (voir par ex. (Siddharthan, 2003)), comme par exemple la formule de Flesch qui combine le nombre de syllabes par mot et le nombre de mots par phrase. Une hypothèse fondamentale de notre approche est qu'une règle de simplification "casse" une structure syntaxique difficile pour la remplacer par une structure syntaxique plus facile à comprendre[10]. Ainsi, la simplification syntaxique effectuée doit corréler avec le nombre de règles appliquées. L'ordonnancement des simplifications utilisé prend donc en compte le nombre d'applications de règles sans doublons dans la sortie[11], puis la taille moyenne des phrases. À la demande de l'utilisateur, une phrase est analysée puis simplifiée, puis les 5 simplifications obtenant les meilleurs scores sont proposées.

Évaluation initiale du système Nous avons développé un jeu de règles destiné à simplifier les textes pour les personnes souffrant d'aphasie. Pour cela, nous avons notamment suivi les résultats expérimentaux obtenus par Caplan (Caplan, 1987) et avons extrait des structures syntaxiques présentant une complexité particulière pour ces patients. Le tableau 1 illustre certains types de phénomènes traités avec un exemple de simplification proposée.

Nous n'avons pas été en mesure de pouvoir nous-mêmes évaluer l'impact sur la compréhension de phrases par des patients aphasiques. L'évaluation de la conservation du sens est un problème très difficile auquel est confronté toute approche de paraphrase. L'applicabilité de notre approche vient du fait que l'auteur du texte d'origine a la possibilité de valider interactivement la simplification proposée par le système qui lui semble la plus adéquate. Il nous reste à évaluer empiriquement l'acceptabilité et l'utilisabilité de notre système au sein d'un traitement de texte.

[9] Sans décision humaine, la simplification obtenant le meilleur score peut être proposée par défaut.

[10] Ceci ne concerne néanmoins pas toutes les règles, puisque certaines ont pour but de réaliser des changements nécessaires pour assurer la cohérence du texte, comme la règle mentionnée plus tôt transformant un groupe nominal indéfini en groupe nominal défini.

[11] Afin de contourner ce problème, nous souhaitons modifier le moteur de simplification pour qu'il applique d'abord les règles qui simplifient les structures les plus "profondes" dans l'arbre syntaxique de la phrase en entrée. Cette approche semble également intuitivement plus proche de la méthodologie de simplification suivie par un humain.

Type de phrase	Exemple de simplification
Passive	The elephant$_i$ was hit t_i by the monkey → The monkey hit the elephant.
Cleft Object	It was the elephant$_i$ [that the monkey hit t_i] → The monkey hit the elephant.
Dative	The elephant gave the rabbit the monkey → The elephant gave the monkey. The rabbit received the monkey.
Conjoined	The elephant [[hit the monkey] and [hugged the rabbit]] → The elephant hit the monkey. The elephant hugged the rabbit.
Subject-Object relative	The elephant$_i$ [that the monkey hit t_i] hugged the rabbit → The monkey hugged the elephant. The elephant hugged the rabbit.
Object-Subject relative	The elephant hit [the monkey that hugged the rabbit] → The monkey hugged the rabbit. The elephant hit the monkey.

Figure 1: Exemples d'application de règles de simplification

3 Perspectives

Un point important pour l'acceptabilité du système concerne le fait que le système doit éviter de demander à l'utilisateur plusieurs validations pour des phrases similaires, et propager des décisions lorsque cela est possible. La validation interactive permet de constituer progressivement un corpus aligné de phrases et de leur simplification, ainsi que d'attribuer des poids aux règles de simplification étant donné l'historique d'une série d'applications de règle. Ces poids pourraient alors être utilisés par le moteur pour guider les simplifications ultérieures d'après la *mémoire de simplification* ainsi constituée.

Remerciements L'auteur remercie David Liben-Nowell pour les nombreuses discussions sur le thème de la simplification syntaxique et pour le travail commun sur le système initial.

Références

Briscoe, Edward et John Carroll (2002), Robust accurate statistical annotation of general text, *Actes de Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Gran Canaria.

Canning, Yvonne (2002), *Syntactic simplification of text*, Thèse de PhD, Université de Sunderland.

Caplan, D. (1987), *Neurolinguistics and Linguistic Aphasiology*, Cambridge University Press, Cambridge.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin et John Tait (1998), Practical Simplification of English Newspaper Text to Assist Aphasic Readers, *Actes de AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Madison, États-Unis.

Chandrasekar, R., C. Doran et B. Srinivas (1996), Motivations and methods for text simplification, *Actes de COLING'96*, Copenhague, Danemark.

Devlin, Siobhan (1999), *Simplifying natural language for aphasic readers*, Thèse de PhD, Université de Sunderland.

Liben-Nowell, David (2000), *Syntactic Simplification*, Thèse de MPhil, Université de Cambridge.

Siddharthan, Advaith (2003), *Syntactic Simplification and Text Cohesion*, Thèse de PhD, Université de Cambridge.