Annotation Linguistique pour l'Évaluation de la Simplification Automatique de Textes*

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter,
Magali Norré, Adeline Müller, Patrick Watrin, Thomas François
CENTAL (IL&C), UCLouvain, Place Montesquieu 3, 1348 Louvain-la-Neuve, Belgique
prenom.nom@uclouvain.be

RÉSUMÉ

L'évaluation des systèmes de simplification automatique de textes (SAT) est une tâche difficile, accomplie à l'aide de métriques automatiques et du jugement humain. Cependant, d'un point de vue linguistique, savoir ce qui est concrètement évalué n'est pas clair. Nous proposons d'annoter un des corpus de référence pour la SAT, ASSET, que nous utilisons pour éclaircir cette question. En plus de la contribution que constitue la ressource annotée, nous montrons comment elle peut être utilisée pour analyser le comportement de SARI, la mesure d'évaluation la plus populaire en SAT. Nous présentons nos conclusions comme une étape pour améliorer les protocoles d'évaluation en SAT à l'avenir.

ABSTRACT

Linguistic Corpus Annotation for Automatic Text Simplification Evaluation

Evaluating automatic text simplification (ATS) systems is a difficult task that is either performed by automatic metrics or user-based evaluations. However, from a linguistic point-of-view, it is not always clear on what bases these evaluations operate. We propose to annotate the ASSET corpus to shed more light on ATS evaluation. In addition to contributing with this resource, we show how it can be used to analyze SARI's relation to linguistic operations. We present our insights as a step to improve ATS evaluation protocols in the future.

MOTS-CLÉS : evaluation, resource, automatic text simplification, annotation.

KEYWORDS: évaluation, ressource, simplification automatique de textes, annotation.

1 Introduction

La simplification automatique de textes (SAT) consiste à rendre des textes plus accessibles pour un public donné. Plusieurs états de l'art sur le sujet ont été publiés ces dernières années (Saggion, 2017; Al-Thanyyan & Azmi, 2021; Štajner, 2021). La SAT est principalement étudiée avec des approches par apprentissage profond (Nisioi *et al.*, 2017; Alva-Manchego *et al.*, 2020b; Cooper & Shardlow, 2020), mais d'autres travaux poursuivent les recherches par le biais de systèmes à base de règles (Evans & Orasan, 2019; Wilkens *et al.*, 2020). Un des verrous majeurs du domaine est l'évaluation des systèmes de SAT (Grabar & Saggion, 2022). Il existe deux approches courantes : le jugement humain et l'évaluation automatique. Dans la première approche, il est demandé à des personnes de

*. Cet article est une adaptation d'un article précédemment publié (Cardon *et al.*, 2022)

noter la sortie d'un système selon trois critères : grammaticalité, préservation du sens, et simplicité. Pour l'évaluation automatique, les métriques les plus courantes sont : BLEU (Papineni *et al.*, 2002), empruntée à la traduction automatique ; SARI (Xu *et al.*, 2016a), spécifiquement proposée pour la SAT ; et la formule de lisibilité Flesch-Kincaid (Kincaid *et al.*, 1975). Bien que ces métriques ne soient pas idéales (Sulem *et al.*, 2018a; Alva-Manchego *et al.*, 2021; Tanprasert & Kauchak, 2021), leur facilité d'utilisation rend leur application répandue. BLEU et SARI nécessitent des simplifications de référence et sont décrites comme étant plus fiables à mesure que le nombre de références augmente. Comme différents publics ont différents besoins en termes de simplification (Rennes *et al.*, 2022), il est crucial de s'assurer qu'un jeu de référence reflète ces besoins. D'autres métriques sont moins fréquemment utilisées, tel que BertScore (Zhang *et al.*, 2020) – initialement conçue pour la génération automatique – et SAMSA (Sulem *et al.*, 2018b) – conçue pour la SAT, mais difficile à utiliser car elle nécessite une annotation sémantique.

En anglais, trois corpus sont couramment utilisés pour l'évaluation : TurkCorpus (Xu *et al.*, 2016b), ASSET (Alva-Manchego *et al.*, 2020a) et Newsela (Xu *et al.*, 2015). Les deux premiers ont les mêmes phrases sources, avec différentes simplifications obtenues par production participative (*crowdsourcing*). Pour les autres langues, qui reçoivent moins d'attention de la communauté, les systèmes sont évalués avec des corpus *adhoc* (Kodaira *et al.*, 2016; Cardon & Grabar, 2020; Anees & Abdul Rauf, 2021; Spring *et al.*, 2021; Todirascu *et al.*, 2022). La manière dont les métriques automatiques sont liées aux opérations de simplification n'est pas connue. Des travaux récents (Vásquez-Rodríguez *et al.*, 2021) ont exploré la relation de ces métriques avec les opérations computationnelles basiques comme l'ajout et la suppression, mais aucun travail n'a étudié cette relation avec les opérations de simplification présentes dans des typologies linguistiques (Brunato *et al.*, 2022; Gala *et al.*, 2020; Amancio & Specia, 2014). Dans cet article, nous souhaitons étudier l'impact du type d'opérations linguistiques sur les métriques automatiques d'évaluation de la SAT. Pour cela, nous avons annoté le corpus ASSET avec les opérations linguistiques qu'il contient.

Nous présentons un état de l'art sur les typologies de transformations pour la simplification (Section 2). Ensuite nous décrivons le processus d'annotation et la ressource produite (Section 3) puis des analyses menées sur SARI avec notre ressource (Section 4). Enfin nous concluons (Section 5) par une synhèse des enseignements obtenus suite à nos expériences.

2 État de l'art

Avant que les méthodes neuronales soient au centre de la recherche en SAT, les typologies d'opérations étaient nécessaires au développement des systèmes, car elles représentaient la base conceptuelle des approches à base de règles. Comme présenté par Siddharthan (2014), les premiers systèmes de SAT s'occupaient de syntaxe (Chandrasekar *et al.*, 1996; Dras, 1999; Brouwers *et al.*, 2014) et décrivaient les typologies d'opérations syntaxiques qui étaient visées pour simplifier la structure des phrases. Actuellement, nous distinguons deux ensembles pour les typologies d'opérations de simplification : l'un basé sur la description linguistique et l'autre basé sur l'édition de chaînes de *tokens*.

2.1 Opérations linguistiques

Le premier ensemble de typologies vise à décrire les opérations linguistiques mises en œuvre lors de la simplification. Cela a été étudié pour différentes langues : l'espagnol (Bott & Saggion, 2014),

l'italien (Brunato *et al.*, 2014, 2022), le français (Koptient *et al.*, 2019; Gala *et al.*, 2020), le portugais du Brésil (Caseli *et al.*, 2009), le basque (Gonzalez-Dios *et al.*, 2018) et l'anglais (Amancio & Specia, 2014). Bien que des opérations soient communes à ces typologies, comme le passage de la voix passive à la voix active, ces typologies présentent des catégories distinctes comme la « *proximization* » (Bott & Saggion, 2014) – faire en sorte que le texte s'adresse au lecteur – ou la « spécification » (Koptient *et al.*, 2019) – conserver un terme difficile et y accoler une explication. Notons que ces deux exemples dépendent du genre de texte : il est peu probable de trouver la *proximization* dans la simplification d'articles encyclopédiques, et il est attendu que la spécification ait lieu dans des textes qui contiennent du lexique spécialisé ou technique. Ces typologies ont été utilisées de manière descriptive, pour renseigner sur les corpus ainsi que sur les pratiques humaines de simplification.

2.2 Éditions de chaînes de tokens

Dans le cadre de ce deuxième ensemble de typologies, les phrases sont vues comme des chaînes de *tokens*, et la simplification consiste à modifier l'agencement de ces *tokens*. Ici, les opérations sont donc décrites comme des modifications à des chaînes de *tokens*, il s'agit par exemple de la suppression, de l'ajout, ou du maintien. Cet angle a été exploré presque exclusivement pour l'anglais (Coster & Kauchak, 2011; Alva-Manchego *et al.*, 2017, 2020a; Vásquez-Rodríguez *et al.*, 2021), avec une exception récente pour l'italien (Brunato *et al.*, 2022). Ces typologies ont été mises en œuvre à diverses fins. Comme pour les typologies linguistiques, elles ont servi à analyser des corpus (Alva-Manchego *et al.*, 2020a). Elles ont aussi servi à étudier la relation des distances entre les chaînes avec les scores attribués par les métriques automatiques (Vásquez-Rodríguez *et al.*, 2021). Certains systèmes de SAT incorporent ce type d'opérations dans leur architecture (Alva-Manchego *et al.*, 2017; Dong *et al.*, 2019; Agrawal *et al.*, 2021). La métrique d'évaluation SARI intègre ce type d'opération dans sa formule : avec les sous-composants KEEP, ADD et DELETE (cf. Section 4 pour plus de détails).

3 Annotation

Deux des trois corpus d'évaluation décrits en section 1 sont librement accessibles : TurkCorpus et ASSET. Nous retenons ASSET, qui a été décrit indépendamment comme meilleur que TurkCorpus (Vásquez-Rodríguez *et al.*, 2021). Cette section présente la typologie d'opérations que nous utilisons pour son annotation (Section 3.1). Nous décrivons le processus d'annotation (Section 3.2), puis nous décrivons la ressource produite (Section 3.3).

3.1 Typologie

Les travaux présentés en section 2.1 proposent des typologies basées sur des analyses manuelles de corpus. Nous construisons la nôtre sur ces travaux. En conséquence, nous n'introduisons aucune nouvelle opération. Nous ne retenons pas les opérations spécifiques au genre de texte, comme celles mentionnées en section 2.1 (*proximization* et spécification), afin d'avoir un ensemble d'opérations génériques. La liste ainsi obtenue est présentée ci-dessous, avec le nom des opérations et leur identifiant dans la ressource. Le nom de certaines opérations est suffisamment descriptif, d'autres opérations sont brièvement clarifiées. Nous présentons les opérations en deux temps : premièrement les opérations qui peuvent directement correspondre à des opérations computationnelles. La correspondance est la

suivante : INSERT (parfois dénotée ADD dans la littérature), DELETE et MOVE, traduits plus bas, sont déjà utilisés tels quels pour les opérations computationnelles. Toutes les autres catégories sont des substitutions.

— **Déplacement** (move)
— **Insérer/Supprimer proposition** (inprop, delprop)
— **Insérer/Supprimer modifieur** (inmod, delmod). Notre définition de modifieur couvre les modifieurs de mots (p. ex. un adjectif qualifiant un nom) et les modifieurs de phrases (p. ex. un complément circonstanciel).
— **Insérer/Supprimer pour la cohérence** (incst, delcst). Toute insertion ou suppression nécessaire suite à une autre opération pour que la phrase reste grammaticale.
— **Insérer/Supprimer autre** (inoth, deloth). Toute insertion ou suppression n'appartenant pas à une des catégories précédentes.
— **Substitution par synonymie** (synonym)
— **Substitution par hyperonymie** (hyperonym)
— **Substitution par hyponymie** (hyponym)
— **Substitution du singulier par le pluriel** (s2p)
— **Substitution du pluriel par le singulier** (p2s)
— **Substitution par pronominalisation** (pron)
— **Substitution par résolution d'antécédent** (fromPron)
— **Modification des traits verbaux** (verbf). Tout changement de mode ou temps d'un verbe.

Deuxièmement, les opérations qui sont le résultat de combinaisons d'opérations computationnelles, ou qui sont trop complexes pour établir une correspondance stable entre les deux types d'opérations.

— **Voix Active vers passive** (a2p)
— **Voix passive vers active** (p2a)
— **Changement de partie du discours** (POSchange)
— **Découpage de phrases** (split)
— **Regroupement de phrases** (merge)
— **Vers forme impersonnelle** (toImp)
— **Vers forme personnelle** (fromImp)
— **Affirmation vers négation** (a2n)
— **Négation vers affirmation** (n2a)

Nous ajoutons également une étiquette **Simplification erronée** (err). Bien que nous n'évaluions pas la simplicité, cela permet de signaler, pendant l'annotation, des erreurs manifestes de grammaticalité ou de préservation du sens, qui rendent la simplification non-désirée en tant que référence dans un protocole d'évaluation.

3.2 Processus

Nous utilisons YAWAT (Germann, 2008) pour l'annotation, un outil déjà utilisé dans ce cadre auparavant (Koptient *et al.*, 2019). [1] La typologie construite fut la base de la rédaction du guide d'annotation. Quatre personnes (travaillant dans la recherche en TAL) ont annoté les mêmes 50 couples de phrases d'ASSET. Cela a servi à (1) évaluer la clarté du guide, (2) former à l'utilisation de l'outil et (3) discuter des points d'amélioration du guide. Cette troisième étape a permis de discuter des cas difficiles et de comment les traiter [2] pour atteindre le consensus. Les discussions n'ont pas

1. Un outil plus récent et commode existe, TS–ANNO (Stodden & Kallmeyer, 2022) mais n'était pas encore disponible au moment de notre annotation.
2. La plupart de ces cas servent d'exemples dans le guide d'annotation.

entraîné de modifications de la typologie. Nous avons réitéré cette étape 2 fois avec 25 nouveaux couples de phrases à chaque fois. Une fois le guide finalisé, nous avons engagé [3] cinq étudiants de master en TAL pour compléter l'équipe d'annotation.

La dernière étape avant l'annotation du corpus intégral – par l'équipe de neuf personnes – fut d'annoter 50 nouveaux couples de phrases d'ASSET. Cela a servi de base au calcul de l'accord inter-annotateurs (Davies & Fleiss, 1982). L'accord est calculé au niveau des tokens et calculé séparément pour les deux côtés du corpus (original et simplifié). L'accord est de 0,61 pour le côté source et de 0,68 pour le côté cible. Nous l'avons également calculé en fusionnant toutes les insertions en une seule catégorie, et toutes les suppressions en une seule catégorie. De la sorte, l'accord est de 0,74 côté source et 0,72 côté cible. Cela indique un compromis entre la granularité de l'annotation et l'accord que l'on peut en obtenir. Le corpus intégralement annoté porte le nom « ASSET$_{ann}$ ».

3.3 Description de la ressource

Cette section décrit le résultat de l'annotation, le corpus ASSET$_{ann}$. Le jeu de test d'ASSET contient 3 590 couples de phrases (359 phrases simplifiées 10 fois chacune). Pendant l'annotation, nous avons trouvé 19 couples de phrases identiques, et 227 simplifications erronées concernant 157 phrases d'origine. ASSET$_{ann}$ contient 3 323 couples de phrases annotées. Un total de 12 827 opérations y sont annotées. *Synonym* est l'opération la plus fréquente, avec 14 % du nombre total d'opérations. Sept opérations (*synonym*, *delcst*, *deloth*, *incst*, *delmod*, *move* et *delprop*) représentent 70 % des opérations à la fois dans le *gold* et dans ASSET$_{ann}$.

4 Analyse de SARI

Nous utilisons ASSET$_{ann}$ pour analyser le comportement de SARI et ses sous-composants en relation avec les opérations de simplification. SARI utilise trois sous-composants, dont la moyenne représente le score final. Ces sous-composants sont *keep*, *add* et *delete*. Pour chaque sous-composant, le score F1 est calculé entre les transformations appliquées de la phrase d'origine pour arriver à la référence ou aux références, et les transformations appliquées de la phrase d'origine pour arriver à la phrase à évaluer. Ces transformations sont observées en n-grammes de *tokens*, où *n* va de 1 à 4 [4] :

$$F1(n, sc) = \frac{2 * prec_{sc}(n) * rappel_{sc}(n)}{prec_{sc}(n) + rappel_{sc}(n)}$$

$$\text{SARI} = \frac{1}{3} \sum_{sc \in \{keep, add, del\}} \frac{1}{k} \sum_{n=1}^{k} F1(n, sc).$$

Nous représentons les couples de phrases par le nombre d'occurrences de chaque opération dans l'annotation. En observant la relation entre la présence d'opérations spécifiques et le score global SARI, nous trouvons une faible corrélation. De plus, même les opérations en correspondance avec les sous-composants de SARI (insertions et suppression) ne sont pas corrélées avec les scores SARI (voir

3. À un taux horaire 25 % supérieur au revenu minimum national.
4. Cette description de SARI correspond à son implémentation dans EASSE (Alva-Manchego *et al.*, 2019), qui est l'outil que nous avons utilisé lors de ce travail.

Table 2 en annexe A), bien qu'elles soient légèrement corrélées avec les scores des sous-composants (voir Table 3 en annexe A).

Pour aller plus loin, nous analysons comment la combinaison des opérations permet de prédire le score SARI, par paire de phrases. Un modèle de régression Lasso (Tibshirani, 1996) avec optimisation des hyperparamètres a ainsi été entraîné et évalué avec R^2, estimé via une validation croisée à 10 plis. Le R^2 de notre modèle est de 0 pour la prédiction du score SARI, ce qui indique que le modèle ne peut pas prédire mieux qu'en utilisant la moyenne. Il ne trouve donc aucun lien entre les opérations de simplification annotées et le score SARI. Nous obtenons le même résultat avec d'autres algorithmes, tels que les arbres de régression (Breiman *et al.*, 1984), les *random forests* (Breiman, 2001) et un perceptron multi-couches (Hinton, 1989) [5].

Cependant, prédire les sous-composants de SARI semble partiellement possible avec Lasso, avec un R^2 moyen de 0,24, 0,03 et 0,23 respectivement pour KEEP, ADD et DEL. La Table 1 présente ainsi les coefficients d'un modèle Lasso entraîné sur le corpus entier pour prédire les sous-composants de SARI. Ces coefficients ont été obtenus avec un R^2 de 0,25, 0,05 et 0,24 pour KEEP, ADD and DEL respectivement. Nous pouvons déjà observer que beaucoup d'opérations ont un coefficient de 0, indiquant qu'elles n'ont pas d'effet sur les sous-composants de SARI.

Il apparaît que bien que SARI montre un certain degré de relation avec les opérations linguistiques, procéder à la moyenne des scores des sous-composants efface cette information. Cela met en lumière deux problèmes de SARI. Premièrement, cette métrique a une variance très faible et n'est pas sensible aux différences entre les sorties des systèmes. Deuxièmement, comme SARI requiert des références, faire la moyenne sur plusieurs références (9 par phrase dans nos expériences) renforce la faible variance. Cette observation est aussi vraie pour les sous-composants, ce qui explique les scores R^2 plutôt faibles.

5 Bilan et perspectives

Nous avons annoté ASSET avec pour objectif d'analyser l'évaluation de la SAT à la lumière des informations linguistiques. Cela nous a permis de porter plusieurs contributions, en plus de la ressource annotée. Concernant l'analyse des pratiques courantes d'évaluation automatique de la SAT, nous avons montré que les sous-composants de SARI peuvent informer sur les opérations linguistiques présentes dans les références et qui apparaissent dans la sortie d'un système, et que cette information est perdue lors du calcul de la moyenne pour produire un score unique. Cela constitue un argument en faveur de rapporter les scores des sous-composants lors de l'évaluation, comme certains travaux commencent à le faire (Zhao *et al.*, 2020; Tanprasert & Kauchak, 2021). Les analyses que nous avons faites encouragent à explorer les liens entre les opérations linguistiques et les opérations computationnelles. Nous pensons qu'ASSET_{ann} et nos expériences représentent une première étape vers des pratiques d'évaluation qui intégreraient ces aspects.

Lors de ce travail, nous avons produit une version annotée du jeu de test d'ASSET. Cette ressource est annotée par 9 personnes, utilisant une typologie que nous proposons en nous basant sur des travaux antérieurs dans plusieurs langues. À partir de l'annotation, nous avons pu nettoyer le jeu de test en excluant 227 couples de phrases avec des erreurs manifestes et produire une version vérifiée manuellement de ces données. Nous avons également mené une analyse poussée de SARI et ses

5. Toutes les expériences ont été menées avec Scikit-learn (Pedregosa *et al.*, 2011)

sous-composants et avons trouvé des liens assez ténus entre ces derniers et les opérations linguistiques. Nous voyons ces résultats comme une direction prometteuse pour l'amélioration de l'évaluation automatique de la SAT, en explorant davantage la relation entre les différents types d'opérations. La ressource décrite ici est librement disponible en ligne. [6]

6 Remerciements

Nous exprimons nos remerciements à Nils Bouckaert, Elena Cao, Angela Kasparian, Melanie Johanns and Luca Matarelli, pour leur aide lors de l'annotation des données. Merci également à Damien de Meyere et Hubert Naets pour leur aide avec YAWAT.

Enfin, nous remercions les relecteurs anonymes pour leurs commentaires et suggestions qui ont contribué à améliorer la qualité de cet article.

Rémi Cardon est financé par le programme *FSR Incoming Postdoc Fellowship* du FSR - Université Catholique de Louvain. Adrien Bibal est financé par la région Wallonne avec un fonds Win2Wal. Rodrigo Wilkens est financé par une convention de recherche avec France Education Internatinal (FEI). David Alfter est financé par le Fonds de la Recherche Scientifique de Belgique (F.R.S-FNRS), référence MIS/PGY F.4518.21.

Références

AGRAWAL S., XU W. & CARPUAT M. (2021). A non-autoregressive edit-based approach to controllable text simplification. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3757–3769, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.330.

AL-THANYYAN S. S. & AZMI A. M. (2021). Automated text simplification : A survey. *ACM Computing Surveys (CSUR)*, **54**(2), 1–36.

ALVA-MANCHEGO F., BINGEL J., PAETZOLD G., SCARTON C. & SPECIA L. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 295–305, Taipei, Taiwan : Asian Federation of Natural Language Processing.

ALVA-MANCHEGO F., MARTIN L., BORDES A., SCARTON C., SAGOT B. & SPECIA L. (2020a). ASSET : A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, p. 4668–4679.

ALVA-MANCHEGO F., MARTIN L., SCARTON C. & SPECIA L. (2019). EASSE : Easier automatic sentence simplification evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing : System Demonstrations*, p. 49–54, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-3009.

ALVA-MANCHEGO F., SCARTON C. & SPECIA L. (2020b). Data-driven sentence simplification : Survey and benchmark. *Computational Linguistics*, **46**(1), 135–187.

6. https://github.com/remicardon/assetann

ALVA-MANCHEGO F., SCARTON C. & SPECIA L. (2021). The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, **47**(4), 861–889.

AMANCIO M. & SPECIA L. (2014). An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, p. 123–130, Gothenburg, Sweden : Association for Computational Linguistics. DOI : 10.3115/v1/W14-1214.

ANEES Y. & ABDUL RAUF S. (2021). Automatic sentence simplification in low resource settings for Urdu. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, p. 60–70, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.nlp4posimpact-1.7.

BOTT S. & SAGGION H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation*, **48**, 93–120.

BREIMAN L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

BREIMAN L., FRIEDMAN J. H., OLSHEN R. A. & STONE C. J. (1984). *Classification and Regression Trees*. Belmont, CA : Wadsworth International Group.

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)@ EACL 2014*, p. 47–56.

BRUNATO D., DELL'ORLETTA F. & VENTURI G. (2022). Linguistically-based comparison of different approaches to building corpora for text simplification : A case study on italian. *Frontiers in Psychology*, **13**.

BRUNATO D., DELL'ORLETTA F., VENTURI G. & MONTEMAGNI S. (2014). Defining an annotation scheme with a view to automatic text simplification. In *Proceedings of the Italian Conference on Computational Linguistics and of the International Workshop EVALITA*, p. 87–92.

CARDON R., BIBAL A., WILKENS R., ALFTER D., NORRÉ M., MÜLLER A., PATRICK W. & FRANÇOIS T. (2022). Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 1842–1866, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

CARDON R. & GRABAR N. (2020). French biomedical text simplification : When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 710–716, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.62.

CASELI H. M., PEREIRA T. F., SPECIA L., PARDO T. A., GASPERIN C. & ALUÍSIO S. M. (2009). Building a brazilian portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science*, **41**, 59–70.

CHANDRASEKAR R., DORAN C. & SRINIVAS B. (1996). Motivations and methods for text simplification. In *The 16th International Conference on Computational Linguistics*.

COOPER M. & SHARDLOW M. (2020). CombiNMT : An exploration into neural text simplification models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 5588–5594, Marseille, France : European Language Resources Association.

COSTER W. & KAUCHAK D. (2011). Simple English Wikipedia : A new text simplification task. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 665–669.

DAVIES M. & FLEISS J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, **38**, 1047.

DONG Y., LI Z., REZAGHOLIZADEH M. & CHEUNG J. C. K. (2019). EditNTS : An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3393–3402, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1331.

DRAS M. (1999). *Tree adjoining grammar and the reluctant paraphrasing of text*. Macquarie University Sydney.

EVANS R. & ORASAN C. (2019). Sentence simplification for semantic role labelling and information extraction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

GALA N., TODIRASCU A., BERNHARD D., WILKENS R. & MEYER J.-P. (2020). Transformations syntaxiques pour une aide à l'apprentissage de la lecture : typologie, adéquation et corpus adaptés. *SHS Web of Conferences*, **78**, 14006.

GERMANN U. (2008). Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL : HLT Demo Session*, p. 20–23, Columbus, Ohio : Association for Computational Linguistics.

GONZALEZ-DIOS I., ARANZABE M. J. & DÍAZ DE ILARRAZA A. (2018). The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation*, **52**(1), 217–247.

GRABAR N. & SAGGION H. (2022). Evaluation of automatic text simplification : Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 453–463, Avignon, France : ATALA.

HINTON G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, **40**(1), 185–234.

KINCAID J., FISHBURNE R., RODGERS R. & CHISSOM B. (1975). *Derivation of new readability formulas for navy enlisted personnel*. Rapport interne, n°8-75, Research Branch Report.

KODAIRA T., KAJIWARA T. & KOMACHI M. (2016). Controlled and balanced dataset for Japanese lexical simplification. In *Proceedings of the ACL Student Research Workshop*, p. 1–7, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-3001.

KOPTIENT A., CARDON R. & GRABAR N. (2019). Simplification-induced transformations : typology and some characteristics. In *Proceedings of the BioNLP Workshop and Shared Task*, p. 309–318.

NISIOI S., ŠTAJNER S., PONZETTO S. P. & DINU L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 2 : Short papers)*, p. 85–91.

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 of the Association for Computational Linguistics*, p. 311–318.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

RENNES E., SANTINI M. & JONSSON A. (2022). The swedish simplification toolkit : – designed with target audiences in mind. In *Proceedings of The 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, p. 31–38, Marseille, France : European Language Resources Association.

SAGGION H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, **10**(1), 1–137.

SIDDHARTHAN A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, **165**(2), 259–298.

SPRING N., RIOS A. & EBLING S. (2021). Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, p. 1339–1349, Held Online : INCOMA Ltd.

ŠTAJNER S. (2021). Automatic text simplification for social good : Progress and challenges. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP*, p. 2637–2652.

STODDEN R. & KALLMEYER L. (2022). TS-ANNO : An annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 145–155, Dublin, Ireland : Association for Computational Linguistics.

SULEM E., ABEND O. & RAPPOPORT A. (2018a). BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 738–744.

SULEM E., ABEND O. & RAPPOPORT A. (2018b). Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 685–696, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1063.

TANPRASERT T. & KAUCHAK D. (2021). Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, p. 1–14, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.gem-1.1.

TIBSHIRANI R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, **58**(1), 267–288.

TODIRASCU A., WILKENS R., ROLIN E., FRANÇOIS T., BERNHARD D. & GALA N. (2022). HECTOR : A hybrid TExt SimplifiCation TOol for raw texts in French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4620–4630, Marseille, France : European Language Resources Association.

VÁSQUEZ-RODRÍGUEZ L., SHARDLOW M., PRZYBYŁA P. & ANANIADOU S. (2021). Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP*, p. 876–882, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.77.

VÁSQUEZ-RODRÍGUEZ L., SHARDLOW M., PRZYBYLA P. & ANANIADOU S. (2021). The role of text simplification operations in evaluation. In *Proceedings of the SEPLN Workshop on Current Trends in Text Simplification*, p. 57–69.

WILKENS R., OBERLE B. & TODIRASCU A. (2020). Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, p. 93–100.

XU W., CALLISON-BURCH C. & NAPOLES C. (2015). Problems in current text simplification research : New data can help. *Transactions of the Association for Computational Linguistics*, **3**, 283–297. DOI : 10.1162/tacl_a_00139.

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

XU W., NAPOLES C., PAVLICK E., CHEN Q. & CALLISON-BURCH C. (2016b). Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, **4**, 401–415. DOI : 10.1162/tacl_a_00107.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating text generation with BERT. In *International Conference on Learning Representations*.

ZHAO Y., CHEN L., CHEN Z. & YU K. (2020). Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(05), 9668–9675. DOI : 10.1609/aaai.v34i05.6515.

A Tableaux d'analyse de SARI

	keep	*add*	*del*
inoth	0	0.1	0
split	0	0.73	0
deloth	-3.91	-0.19	3.05
delcst	-2.44	0	1.52
move	-1.55	0.14	1.5
delprop	-3.97	-0.74	3.66
incst	0	0.95	0
hyperonym	0	0.14	1.58
synonym	-1.68	0.61	3.63
delmod	-3.39	0	3.02

TABLE 1 – Coefficients des modèles de régression Lasso pour la prédiction des sous-composants de SARI. Les opérations ayant des coefficients différents de zéro pour *add* et *del* impliquent l'ajout et la suppression de *tokens*. Les coefficients négatifs pour la prédiction de *keep* indiquent que l'opération diminue le score du sous-composant *keep* de SARI. Les opérations absentes ont tous leurs coefficients à zéro.

	Spearman correlation	p-value
inoth	0.0161	0.3349
split	0.0915	0
deloth	-0.0137	0.4119
p2s	-0.017	0.3084
delcst	0.0413	0.0134
verbf	0.0519	0.0018
pron	-0.035	0.036
move	0.012	0.4731
inprop	0.0295	0.0775
delprop	-0.0381	0.0223
incst	0.0718	0
s2p	-0.0338	0.0428
fromPron	-0.0038	0.8194
merge	-0.0033	0.8421
p2a	0.0119	0.4752
pos2neg	0.0028	0.8648
neg2pos	-0.018	0.2806
hyponym	-0.0191	0.253
toImp	-0.0283	0.09
fromImp	-0.0349	0.0367
POSchange	0.0177	0.2895
hyperonym	0.0374	0.025
a2p	-0.018	0.2814
synonym	0.163	0
delmod	-0.0025	0.8792
inmod	0.0194	0.2451

TABLE 2 – Corrélation de Spearman entre l'occurrence des opérations dans les couples de phrases et le score SARI. Une p-value marquée à 0 signifie qu'elle est inférieure à 0,0001. Les p-values élevées s'expliquent par un nombre insuffisant d'occurrences des opérations correspondantes dans le corpus.

	keep		add		del	
Transformation	Spearman	p-value	Spearman	p-value	Spearman	p-value
inoth	-0.0752	0	0.0658	0.0001	0.073	0
split	0.0251	0.1328	0.1925	0	-0.0063	0.7052
deloth	-0.2214	0	0.0008	0.9614	0.2015	0
p2s	-0.0478	0.0042	0.0144	0.3899	0.0378	0.0235
delcst	-0.2267	0	0.1482	0	0.2279	0
verbf	-0.0827	0	0.089	0	0.1441	0
pron	-0.0827	0	0.011	0.5084	0.0321	0.0547
move	-0.1764	0	0.0828	0	0.1486	0
inprop	-0.0699	0	0.069	0	0.0947	0
delprop	-0.1694	0	-0.0636	0.0001	0.1809	0
incst	-0.1269	0	0.2198	0	0.1362	0
s2p	-0.108	0	0.0258	0.1217	0.073	0
fromPron	-0.0458	0.006	0.0114	0.4931	0.04	0.0165
merge	-0.0257	0.123	-0.0056	0.736	0.0269	0.107
p2a	-0.0304	0.0684	0.0123	0.4616	0.0421	0.0116
pos2neg	-0.038	0.0229	0.0186	0.2663	0.0346	0.038
neg2pos	-0.038	0.0226	-0.0147	0.3777	0.0378	0.0235
hyponym	-0.0373	0.0256	0.0263	0.1145	0.0068	0.6847
toImp	-0.0826	0	0.0148	0.3767	0.049	0.0033
fromImp	-0.0355	0.0333	-0.0278	0.0957	-0.0055	0.7408
POSchange	-0.1317	0	0.0655	0.0001	0.1538	0
hyperonym	-0.0303	0.0699	0.0836	0	0.0649	0.0001
a2p	-0.0307	0.0658	-0.0281	0.0926	0.0155	0.353
synonym	-0.0607	0.0003	0.2135	0	0.2116	0
delmod	-0.1857	0	0.0195	0.2416	0.2071	0
inmod	-0.0472	0.0047	0.0332	0.0466	0.0725	0

TABLE 3 – Corrélations de Spearman et p-values entre chaque transformation annotée et les sous-composant de SARI. Une p-value marquée à 0 signifie qu'elle est inférieure à 0,0001.