Overview of the SimpleText Workshop at INFORSID-2021: Scientific Text Simplification and Popularization

Liana Ermakova [1], Josiane Mothe [2] and Eric Sanjuan [3]

[1] *HCTI - EA 4249, Université de Bretagne Occidentale, 20 Rue Duquesne, 29200 Brest, France*
[2] *INSPE, Université de Toulouse, IRIT, UMR5505 CNRS, Toulouse, France*
[3] *Avignon Université, LIA, Avignon, France*

Abstract

Although scientific literature is accessible for citizens, it remains difficult to read for non-experts because of its linguistic complexity and structure and lack of background knowledge for readers. Text simplification aims to reduce these obstacles. In this paper we present an overview of the SimpleText workshop at the French conference INFORSID-2021. The INFORSID'21 SimpleText Workshop addresses the opportunities and challenges of scientific text simplification approaches to improve access to information and scientific acculturation. The INFORSID'21 SimpleText Workshop relies on an interdisciplinary community of researchers in automatic language processing, information retrieval, linguistics, sociology, science journalism and science popularization working together to try to solve one of the biggest challenges of today. Five scientific papers covering the fields of medical text simplification, didactics, translation and proofreading, artificial intelligence and technical writing have been selected for publication.

Keywords

Scientific text simplification, science popularization, workshop overview

1. Introduction

Key information from authoritative primary sources is available to citizens via modern information access systems. In reality, understanding of scientific literature remains difficult for non-experts because of its linguistic complexity, structure, length of scientific documents and lack of background scientific knowledge for readers in general.

The content of the scientific debate which proceeds by confronting a multiplicity of studies before reaching a consensus adds complexity. Individual or political decisions are potentially impacted by a lack of knowledge of all current scientific work and debates. Scientists may also be confronted with reading difficulty when they are interested in scientific documents from disciplines in which they are not experts. Contradictory results within a discipline are difficult for non-specialists to understand. The situation is even worse in case of potentially contradictory results in different disciplines.

Text simplification aims to reduce some of these obstacles. Thus, the INFORSID'21 SimpleText Workshop addresses the opportunities and challenges of scientific text simplification approaches to improve access to information and provide relevant background knowledge. SimpleText aims to take a step towards truly open, accessible and understandable science for everybody, help counteract fake news based on scientific results; as well as enable faster reading, which can also facilitate access to scientific results. This is particularly important given the soaring of open science and preprints during the COVID-19 pandemic (e.g. SAGE Publishing[2] and Springer Nature[3] have made COVID-19 publications open access). More than 1800 computer-science papers deal with "covid" in the *arXiv*

[1]CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
EMAIL: liana.ermakova@univ-brest.fr (A. 1); josiane.mothe@irit.fr (A. 2); eric.sanjuan@univ-avignon.fr (A. 3)
ORCID: 0000-0002-7598-7474 (A. 1)

CEUR Workshop Proceedings (CEUR-WS.org)

[2] https://journals.sagepub.com/coronavirus
[3] https://www.springernature.com/fr/researchers/campaigns/coronavirus

computer science repository in the last 12 months[4]. Simplified texts may also be intended to be more accessible to non-native speakers, young readers, and people with reading disabilities. Text simplification can improve natural language processing applications, including machine translation outputs.

Automatic text simplification could be useful in various fields such as scientific communication, science journalism, politics, computer-assisted translation, technical writing, and education.

The INFORSID'21 SimpleText Workshop is a place to think about and work on these issues. This workshop is part of MaDICS[5] (Masses de Données, Informations et Connaissances en Sciences - Big data, Information and Knowledge in Sciences), a French national research group on Big Data - Data Science. SimpleText was accepted as a workshop at the French conference INFORSID-2021[6] (INFormatique des ORganisations et Systèmes d'Information et de Décision - Computer science for organisations, information and decision systems).

2. Selected papers

Five scientific papers covering the fields of medical text simplification, didactics, translation and proofreading, artificial intelligence and technical writing have been selected for publication in the proceedings "INFORSID Workshops - Designing the future of information systems together" (in alphabetical order) [1]:

- Sílvia Araújo, Radia Hannachi “Pour une démarche de communication multimodale de données scientifiques : de la recherche documentaire à l'infographie via le mind mapping” (“For a multimodal communication of scientific data: from documentary research to computer graphics via mind mapping”) [2];
- Rémi Cardon, Natalia Grabar “Recherche de phrases parallèles à partir de corpus comparables pour la simplification de textes médicaux en français” (“Search for parallel sentences from comparable corpora for the simplification of medical texts in French”) [3];
- Helen Mccombie-Boudry “Could automatic text simplification assist correction-revision of scientific texts written by non-native English speakers?” [4];
- John Rochford “Developing Simple Web Text for People with Intellectual Disabilities and to Train Artificial Intelligence” [5];
- Mike Unwalla “Controlled language for text simplification: Concepts and implementation” [6].

Our speakers Radhia Hannachi (HCTI, Université de Bretagne Sud) and Sílvia Araújo (Research Team on Digital Humanities, Universidade do Minho, Portugal) are specialist in didactics and presented the talk “For a multimodal communication of scientific data: from documentary research to computer graphics via mind mapping”.

Two guest speakers presented their work at SimpleText@INFORSID: J. Rochford (US) and N. Grabar (EPST - NLP, UMR 8163).

J. Rochford’s team is training AI to simplify web text (EasyText.AI) and participates in the EasyCOVID-19 project. The EasyCOVID-19 project aims to simplify textual information from every world’s government websites but it does not tackle scientific literature. Simplified texts are also more accessible for non-native speakers [7], young readers, people with reading disabilities [8], [9] or lower levels of education (sustainable development goal REDUCED INEQUALITY).

Grabar and Cardon introduced a corpus of technical and simplified medical texts in French [10], [11]. The corpus contains 663 pairs of comparable sentences retrieved from encyclopedias, medication information leaflets and scientific summaries, and aligned by two annotators. In [11], they proposed an automatic method for sentence alignment. In a further work, they trained neural models on (1) the comparable health corpus in French, (2) the WikiLarge corpus translated from English to French, and (3) and a lexicon that associates medical terms with paraphrases, using different ratios of general and specialized sentences [12]. Jiang et al. proposed a neural CRF alignment model and constructed two text simplification datasets: Newsela-Auto and Wiki-Auto [13]. Their transformer-based seq2seq model established a new state-of-the-art text simplification method in both automatic and human evaluation.

[4] https://arxiv.org/search/advanced?advanced=&terms-0-operator=AND&terms-0-term=covid&terms-0-field=all&classification-computer_science=y&classification-physics_archives=all&classification-include_cross_list=include&date-filter_by=past_12&date-year=&date-from_date=&date-to_date=&date-date_type=submitted_date&abstracts=show&size=50&order=-announced_date_first

[5] https://www.madics.fr/ateliers/simpletext/

[6] https://inforsid2021.sciencesconf.org/

As stated in [11], no parallel data (simplified and not) in French exists. Besides, their work tackles language simplification only without considering content selection for popularized texts which can be different from those designed for experts. N. Grabar accepted to be an invited speaker at SimpleText@INFORSID in 2021.

Many organizations with a multicultural environment or organizations targeting international audiences use plain languages, for example, ‘plain English’ and ‘lenguaje claro’ (plain Spanish). However, for safety-critical documentation, plain language is not always sufficient, and some organizations use controlled language, e.g. the ASD-STE100 Simplified Technical English specification. According to ISO/TS 24620-1:2015, a controlled language is a “subset of natural languages whose grammars and dictionaries have been restricted in order to reduce or eliminate both ambiguity and complexity”. Controlled languages are widely used in the technical documentation industry but they are also applicable to other texts. In technical texts, phrasal verbs may cause problems for non-native readers (see ASD-STE100 rule 9.3: when you use two words together, do not make phrasal verbs), while synonyms can be confusing. Controlled language specifications, like ASD-STE100, have (1) a dictionary of approved and unapproved terms and (2) a set of writing rules and (3) require terminology consistency (e.g. ASD-STE100 rule 1.11: do not use different technical names for the same thing). There are several solutions to check a document for compliance to the ASD-STE100 specification: Boeing Simplified English Checker (BSEC), HyperSTE, TermChecker. Some work was carried out for Spanish [62], but to the best of our knowledge no validation parser for controlled languages in French exists. Many of the ASD-STE100 rules are applicable to the simplification of scientific texts, but, as far as we know, no work has been carried out on application of these rules to scientific text simplification. M. Unwalla (UK) presented an industrial talk at SimpleText@INFORSID in 2021 on the TechScribe term checker, which checks a document for compliance to the ASD-STE100 Simplified Technical English specification for a controlled language.

3. Conclusions and discussions

The INFORSID'21 SimpleText Workshop relies on an interdisciplinary community of researchers in automatic language processing, information retrieval, linguistics, sociology, science journalism and science popularization working together to try to solve one of the biggest challenges of today.

Text simplification is widely used in the field of translation and localization. Localization requires internalization, i.e. preparation of computer software, websites etc. by isolation of textual content, language simplification and eliminating culture-specific data. From the linguistic point of view, internalization is closely related to pre-editing. Pre-editing is the process whereby a human prepares a document before applying machine translation to achieve better results [14] and to decrease post-editing workload which has become a mainstream choice for companies publishing their content in different languages. Text simplification can be used for proof-reading, for example, in 2021, the UBO translation office will give a talk at SimpleText@INFORSID “Could automatic text simplification assist correction-revision of scientific texts written by non-native English speakers?”.

Scientific literacy, including health related questions, is important in order to allow people to make informed decisions, evaluate information quality, maintain physiological and mental health. For example, the stories individuals find believable can determine their response to the COVID-19 pandemic, including the application of social distancing, using dangerous fake medical treatments, or hoarding panic buying. Unfortunately, stories in social media are easier for non-experts to understand than research papers due to the lack of prior background knowledge or complex language and internal vernacular. Scientific texts such as research publications can also be difficult to understand for non field experts or scientists outside the publication field. From a societal perspective, SimpleText is a step forward to make research really open and accessible for everyone [15], to develop a counter-speech to fake news based on scientific results, to allow people to read faster and consequently, become more aware of scientific results (sustainable development goal QUALITY EDUCATION). However, improving text clarity and its adaptation to different audiences remains an unsolved problem.

4. Acknowledgements

The publications have been reviewed by the interdisciplinary program committee from various fields such as computer science, linguistics, epistemology and history of science and technology. We would like to thank our program committee:

- Ismail Badache, Univ. d’Aix-Marseille, Univ. de Toulon CNRS, LIS UMR 7020 (Marseille, France)
- Patrice Bellot, Univ. d’Aix-Marseille, Univ. de Toulon CNRS, LIS UMR 7020 (Marseille, France);
- Pavel Braslavski, Combinatorial Algebra Lab, Ural Federal University (Yekaterinburg, Russia);
- Adrian-Gabriel Chifu, Univ. d’Aix-Marseille, Univ. de Toulon CNRS, LIS UMR 7020 (Marseille, France);
- Liana Ermakova, HCTI - EA 4249, Université de Bretagne Occidentale (Brest, France);
- Hervé Ferrière, Centre François Viète d'épistémologie et d'histoire des sciences et des techniques, Université de Bretagne Occidentale (Brest, France);
- Jaap Kamps, Faculty of Humanities, University of Amsterdam (Amsterdam, Netherland);
- Élise Mathurin, HCTI - EA 4249, Université de Bretagne Occidentale (Brest, France);
- Josiane Mothe, INSPE, Université de Toulouse, IRIT, UMR5505 CNRS (Toulouse, France);
- Diana Nurbakova, LIRIS, Institut National des Sciences Appliquées de Lyon, (Lyon, France);
- Irina Ovchinnikova, Institute of Linguistics and Intercultural Communication, Sechenov University (Moscow, Russia);
- Michael Rinn, HCTI - EA 4249, Université de Bretagne Occidentale (Brest, France);
- Eric San-Juan, Laboratoire d’Informatique d’Avignon, Institut de technologie d' Avignon (Avignon, France).

.

5. References

[1] L. Ermakova, J. Mothe, and E. Sanjuan, ‘Atelier SimpleText : Simplification et Vulgarisation des Textes Scientifiques - Préface’, in *Actes des Ateliers d’INFORSID - Dessinons ensemble le futur des systèmes d’information*, 2021, pp. 57–60. [Online]. Available: http://inforsid.fr/actes/2021/ActesAteliers_INFORSID2021.pdf#page=63

[2] S. Araújo and R. Hannachi, ‘Pour une démarche de communication multimodale de données scientifiques : de la recherche documentaire à l’infographie via le mind mapping’, in *Actes des Ateliers d’INFORSID - Dessinons ensemble le futur des systèmes d’information*, 2021, pp. 64–74. [Online]. Available: http://inforsid.fr/actes/2021/ActesAteliers_INFORSID2021.pdf#page=63

[3] R. Cardon and N. Grabar, ‘Recherche de phrases parallèles à partir de corpus comparables pour la simplification de textes médicaux en français’, in *Actes des Ateliers d’INFORSID - Dessinons ensemble le futur des systèmes d’information*, 2021, pp. 61–63. [Online]. Available: http://inforsid.fr/actes/2021/ActesAteliers_INFORSID2021.pdf#page=63

[4] H. McCombie, ‘Could automatic text simplification assist correction-revision of scientific texts written by non-native English speakers?’, in *Actes des Ateliers d’INFORSID - Dessinons ensemble le futur des systèmes d’information*, 2021, pp. 80–87. [Online]. Available: http://inforsid.fr/actes/2021/ActesAteliers_INFORSID2021.pdf#page=63

[5] J. Rochford, ‘Developing Simple Web Text for People with Intellectual Disabilities and to Train Artificial Intelligence’, in *Actes des Ateliers d’INFORSID - Dessinons ensemble le futur des systèmes d’information*, 2021, pp. 88–95. [Online]. Available: http://inforsid.fr/actes/2021/ActesAteliers_INFORSID2021.pdf#page=63

[6] M. Unwalla, ‘Controlled language for text simplification: Concepts and implementation’, in *Actes des Ateliers d’INFORSID - Dessinons ensemble le futur des systèmes d’information*, 2021, pp. 75–79. [Online]. Available: http://inforsid.fr/actes/2021/ActesAteliers_INFORSID2021.pdf#page=63

[7] A. Siddharthan, ‘An architecture for a text simplification system’, presented at the Proceedings of the Language Engineering Conference 2002 (LEC 2002), 2002. Accessed: Nov. 20, 2020. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.9968&rank=1

[8] P. Chen, J. Rochford, D. N. Kennedy, S. Djamasbi, P. Fay, and W. Scott, ‘Automatic Text Simplification for People with Intellectual Disabilities’, in *Artificial Intelligence Science and Technology*, 0 vols, WORLD SCIENTIFIC, 2016, pp. 725–731. doi: 10.1142/9789813206823_0091.

[9] K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura, ‘Text simplification for reading assistance: a project note’, in *Proceedings of the second international workshop on Paraphrasing - Volume 16*, USA, Jul. 2003, pp. 9–16. doi: 10.3115/1118984.1118986.

[10] N. Grabar and R. Cardon, ‘CLEAR-Simple Corpus for Medical French’, presented at the ATA, Nov. 2018. Accessed: Apr. 22, 2021. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-01968355

[11] R. Cardon and N. Grabar, ‘Détection automatique de phrases parallèles dans un corpus biomédical comparable technique/simplifié’, Toulouse, France, Jul. 2019. Accessed: Apr. 22, 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02430446

[12] R. Cardon and N. Grabar, ‘French Biomedical Text Simplification: When Small and Precise Helps’, in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), Dec. 2020, pp. 710–716. doi: 10.18653/v1/2020.coling-main.62.

[13] C. Jiang, M. Maddela, W. Lan, Y. Zhong, and W. Xu, ‘Neural CRF Model for Sentence Alignment in Text Simplification’, *ArXiv200502324 Cs*, Jun. 2020, Accessed: Apr. 23, 2021. [Online]. Available: http://arxiv.org/abs/2005.02324

[14] P. Bouillon, L. Gaspar, J. Gerlach, V. Porro, and J. Roturier, ‘Pre-editing by Forum Users: a Case Study’, p. 8.

[15] B. Fecher and S. Friesike, ‘Open science: one term, five schools of thought’, in *Opening science*, Springer, Cham, 2014, pp. 17–47.