Bridging the Gap Between Qualitative and Quantitative – How Linguistic Analysis Can Help Automatic Text Simplification Evaluation

Noémi Prótár
Eötvös Loránd University
Budapest, Hungary
protarnoemi@student.elte.hu

Dávid Márk Nemeskey
Department of Digital Humanities
Eötvös Loránd University
Budapest, Hungary
nemeskey.david@btk.elte.hu

Abstract

Automatic text simplification has long faced the problem of automatic evaluation: the scarcity of good-quality reference data and the differences between languages make both n-gram based and deep learning-based scoring systems unreliable.

Our research explores how quantitative, automated scoring methods can be integrated with insights from qualitative linguistic analysis. The aim is to enhance the ability of automated systems to reflect the structural and content-related aspects of texts, while also improving their suitability for evaluating simplified Hungarian texts.

By aligning surface-level metrics with deeper linguistic features, our approach addresses a critical gap in current evaluation practices. While focused on Hungarian text simplification, the methodology may also contribute to more robust evaluation frameworks for other low-resource or morphologically rich languages.[1]

Introduction

Text simplification is a prevalent research topic in Natural Language Processing: as it is concerned with making textual data more accessible for a wider range of people, it poses not only scientifically relevant questions, but it also tries to solve a social need. As text simplification for humans can be time and energy consuming, and therefore very expensive, it has long been an interest of natural language processing to facilitate this by automating text simplification. These automatically generated texts, however, have to be evaluated. Researchers usually do not have the means to hire and train professional annotators for this task. Therefore the need for reliable automatic evaluation tools or metrics is high.

In our research we will explore the possibilities of combining quantitative, automatic scoring methods with the conclusions drawn from qualitative linguistic analyses. Our goal is to make automatic scoring systems reflect the overall structure and content of the text better, as well as making automatic scoring systems better suited for Hungarian.

Research Question

This study aims to explore how qualitative and quantitative methods can be used in combination, in order to optimize automatic evaluations for text simplification. Therefore the research poses the following questions:

- What linguistic patterns emerge from the comparative analysis of Hungarian standard-language and simplified texts? What are the linguistic characteristics of simpified texts?
- How can linguistic analysis help aligning automatic scoring systems for Hungarian with the expected output?

[1] The authors acknowledge the support of the National Laboratory for Digital Heritage. Project no. 2022-2.1.1-NL-2022-00009 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2022-2.1.1-NL funding scheme.

Methodology

Qualitative analysis can be supported by quantitative analysis and vice versa. By themselves, both paradigms face challenges: qualitative analysis can only be carried out on a few data points, due to their great demand of human resource, while quantitative analysis cannot take individual data points into consideration. A multifaceted approach may alleviate some of these limitations.

Our goal in this research is to bridge the gap between qualitative and quantitative analyses regarding automatic scoring in text simplification. For this we will be using HunSimpleNews [Prótár and Nemeskey, 2025], which is the only currently available document-level text simplification corpus for Hungarian. As the pairings' correctness in the corpus was verified by annotators, the corpus offers a good baseline for these experiments.

There are multiple approaches to automatically evaluate an automatically generated simplified text. Some of them are only suited for sentence-level simplifications: EASSE (Easier Automatic Sentence Simplification Evaluation) [Alva-Manchego et al., 2019] collects the most important sentence simplification metrics. It is able to measure for example the SARI-score [Xu et al., 2016] and the BLEU-score [Papineni et al., 2002] that are the most commonly used metrics to measure if a text simplification is adequate, both on sentence and on document-level. These scores are, however, optimized for non-agglutinative languages, as they are n-gram based, therefore for Hungarian they might report a lower score. They also only partly account for the variety in language: SARI is able to handle multiple reference texts, but most datasets simply do not have multiple reference texts to begin with. It also features SAMSA [Sulem et al., 2018b], which is a simplicity-specific metric that focuses on sentence splitting and the Flesch Kincaid Grade Level (FKGL), which is a readability metric, but which – according to the authors – needs to be interpreted with caution, as it rewards agrammatical sentences if they are short and consist of short words. They also provide some test datasets such as PWKP [Zhu et al., 2010], TurkCorpus [Xu et al., 2016] and HSplit [Sulem et al., 2018a]. All of these corpora are, however, in English, therefore they cannot be used for Hungarian evaluation.

A significant portion of text simplification research does not focus on syntactic simplification, but on lexical simplification: as Bredel and Maaß [2016] and Maaß [2015] underline, simplified texts need to use central elements from the lexicon: as central elements are used more often, they are more conventional and so they require less mental work to understand and to use (cf. Tolcsvai Nagy [2017]). Therefore in our analysis we have to consider not only the document- and sentence-level alterations, but also the overall vocabulary the simplified text uses.

In our research we will examine how these scoring methods can be better adapted for agglutinative languages, such as Hungarian. In our research we will carry out a qualitative, linguistics-based analysis on a selected few text pairs. In the analysis we will concentrate on

- contextualization strategies in the text (cf. Imrényi [2017] and Tátrai [2017]),
- the information structure of the text, i.e. what information is omitted or added, and in what order,
- how simplified and standard language texts use quotations (cf. Csontos [2023]).

As the simplified texts presumably use different strategies than standard-language texts do in these realms, it can be hypothesized that the construal strategies can serve as a baseline for evaluation. Therefore we will then evaluate the qualitative findings, and collect emergent schemata, i.e. phenomena that occur reliably in multiple texts. Then we will examine how these findings can be translated to qualitative, automatically evaluable attributes.

We will also carry out quantitative analyses regarding the POS-tag distribution, in which we will compare the simplified and the original texts. This will help us to find the most important structures that are commonly used in simplified texts. We will also carry out a semantic analysis, in which we will compare the words used in both domains to the most used words of a large reference corpus, the Hungarian Webcorpus [Halácsy et al., 2004, Kornai et al., 2006], presuming that the most used words in the reference corpus are central elements in the Hungarian lexicon, and therefore more conventional, thus easier to use. These analyses will then be translated to an evaluation system using the findings.

As Large Language Models are capable of evaluation, we will also construct an LLM-as-a-judge system, to see if a model is able to identify weak points in automatically generated texts.

Preliminary Findings

The preliminary linguistic experiments on a small number of texts have shown that there are considerable qualitative differences between the source and target texts. [Anonymized, 2025] A qualitative study was carried out regarding contextualization, which proved that simplified texts operate with vastly different contextualization strategies than standard-language texts. As there are significant, observable differences between the two domains, the schemata found in the simplified texts can be used to evaluate whether a text is properly simplified or not, based on the assumption that if a text is similar to simplified texts, then it is a simplified text itself.

Preliminary findings in a different research question were presented at the XXVIII. Spring Wind conference, in June 2025. This pilot study has already tried to combine both qualitative and quantitative methods, mainly in prompt tuning. The study indicates that corrections in the prompt based on results of qualitative studies can increase the alignment of the LLM, even though the quality improvement may not translate directly to automatic scoring systems: the scores get higher with each iteration, but they are always within the range of standard deviation. This, however, has indicated for us, that conclusions drawn based on qualitative analyses can help successful simplified text generation, and it has also shown us that the current automatic scoring systems cannot adequately measure whether a simplified text's quality has improved.

Impact

Our research's goal is to provide an automatic evaluation tool for Hungarian simplified texts. This research will advance our understanding of the internal structure of simplified texts and offer a reliable, automated method for evaluating Hungarian simplified texts. Although our focus is Hungarian, the methodology we propose may inspire multilingual simplification evaluation frameworks, especially for low-resource or morphologically rich languages.

References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. EASSE: Easier automatic sentence simplification evaluation. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3009. URL `https://aclanthology.org/D19-3009/`.

Anonymized. Anonymized. In *Anonymized*, 2025.

Ursula Bredel and Christiane Maaß. *Ratgeber Leichte Sprache*. Duden, Berlin, 2016.

Nóra Csontos. *Az idézés működése a magyar nyelvben. Funkcionális kognitív közelítés*. Károli könyvek. Monográfia. L'Harmattan, Budapest, 2023.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL `https://aclanthology.org/L04-1320/`.

András Imrényi. Mondattan. In Gábor Tolcsvai Nagy, editor, *Nyelvtan*, chapter Pragmatika, pages 663–760. Osiris Kiadó, Budapest, Hungary, 2017.

András Kornai, Péter Halácsy, Viktor Nagy, Csaba Oravecz, Viktor Trón, and Dániel Varga. Web-based frequency dictionaries for medium density languages. In *Proceedings of the 2nd International Workshop on Web as Corpus*, 2006. URL https://aclanthology.org/W06-1701/.

Christiane Maaß. *Leichte Sprache. Das Regelbuch.* Lit, Münster, 2015. doi: 10.25528/018. URL https://doi.org/10.25528/018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Noémi Prótár and Dávid Márk Nemeskey. HunSimpleNews: Az első autentikus magyar nyelvű szövegekből álló szövegegyszerűsítési korpusz. In *XXI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2025)*, pages 197–218, Szeged, 2025.

Elior Sulem, Omri Abend, and Ari Rappoport. BLEU is not suitable for the evaluation of text simplification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1081. URL https://aclanthology.org/D18-1081/.

Elior Sulem, Omri Abend, and Ari Rappoport. Semantic structural evaluation for text simplification. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-1063. URL https://aclanthology.org/N18-1063/.

Gábor Tolcsvai Nagy. Bevezetés. In Gábor Tolcsvai Nagy, editor, *Nyelvtan*, chapter Pragmatika, pages 23–71. Osiris Kiadó, Budapest, Hungary, 2017.

Szilárd Tátrai. Pragmatika. In Gábor Tolcsvai Nagy, editor, *Nyelvtan*, chapter Pragmatika, pages 899–1058. Osiris Kiadó, Budapest, Hungary, 2017.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. volume 4, pages 401–415, 2016. URL https://www.aclweb.org/anthology/Q16-1029.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. volume 2, pages 1353–1361, 08 2010.