Automatic Literary Adaptation? A Survey of Related Tasks, Trends and Challenges

Iglika Nikolova-Stoupak
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
iglika.nikolova-stoupak@etu.sorbonne-universite.fr

Gaël Lejeune
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
gael.lejeune@sorbonne-universite.fr

Eva Schaeffer-Lacroix
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
eva.lacroix@inspe-paris.fr

This paper serves as a survey of the contemporary Natural Language Processing (NLP) landscape with regard to its potential to perform or assist in literary adaptation - the modification of literature in order to meet the needs and preferences of specific reader audiences, such as children, foreign language learners and people with learning disabilities. Firstly, we present the need for and nature of literary adaptation. Then, we provide overviews of the current contexts in relation to existing NLP tasks that are of direct relevance to literary adaptation: simplification, summarisation, style transfer, and creative text generation. Throughout the survey, we focus on large language models (LLMs) as the current state of the art, as well as on existing methods of evaluation of automatically generated text and, where possible, on multilingual applications. We wrap up with a discussion of future directions for literary adaptation as the intersection of the discussed tasks. In our view, contemporary NLP has the potential to engage in this highly novel task, given an established multidimensional evaluation framework and ongoing engagement on the side of professionals in fields such as education and creative writing.

Keywords: simplification, summarisation, style transfer, creative generation, survey, literary adaptation

1. Introduction

To date, Natural Language Processing (NLP) has not consistently engaged in the generation of adapted literary text that aims to meet the needs of specific readerships (such as children of a defined age, language learners of a defined level, or people with learning disabilities). Whilst related tasks such as text simplification and endeavours like easy-read (Freyer, Kempt, and Klöser 2024) come close in terms of definition and purpose and do not intrinsically exclude application to literature, the current focus remains on information texts. Therefore, no unified framework and practices are established regarding the distinguishing

https://doi.org/10.47810/JCLIB.2.2026.04

features of literature, such as large textual length, narrative voice, style, and figurative language.

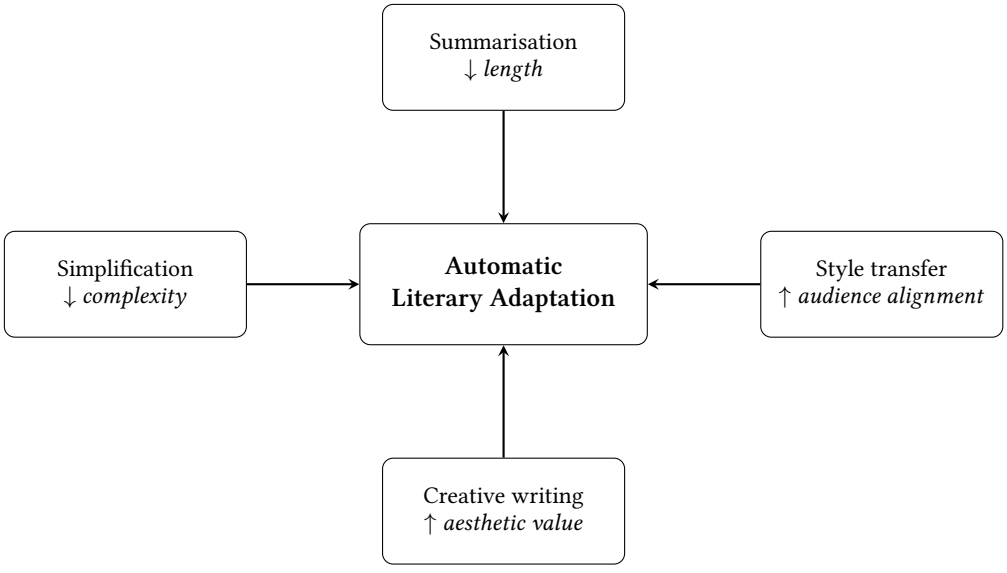

Figure 1
Automatic literary adaptation at the crossroads of relevant NLP tasks.

In this survey-style paper, we lay out the current landscapes in relation to NLP (more specifically, "natural language generation", or "NLG") tasks that are of direct relevance to the concept of automatic literary adaptation: simplification, summarisation, style transfer, and creative generation. See Fig. 1 for an overview of these tasks' significance to the topic at hand. We performed a bibliometric analysis in order to summarise and visualise a timeline of the interest in the mentioned tasks through the proxy of the related articles in the platform Semantic Scholar[1]. We utilised Semantic Scholar's official API to determine the number of title matches per year (1950 to 2026) of search terms elaborated to represent each task[2]. We also limited the results by the domain "Computer Studies". As Fig. 2 shows, interest (and, coincidentally, advancement) in all four tasks has been growing through the years. Summarisation is represented by the highest number of articles overall as well as by significant growth since the early 2000s i.e. the time when statistical models became largely used. The other three tasks, for which the aspect of semantics has a stronger role, are associated with peaks in the late 2010s and early 2020s, when LLMs were introduced.

1 https://www.semanticscholar.org/
2 simplification: SIMPLIFICATION OR SIMPLIFY OR SIMPLIFYING
summarisation: SUMMARIZATION OR SUMMARISATION OR SUMMARIZING OR SUMMARISING
style transfer: STYLE TRANSFER OR STYLE ADAPTATION
creative generation: CREATIVE WRITING OR CREATIVE GENERATION OR STORY GENERATION OR STORY WRITING OR NARRATIVE GENERATION OR NARRATIVE WRITING OR POETRY GENERATION OR POETRY WRITING

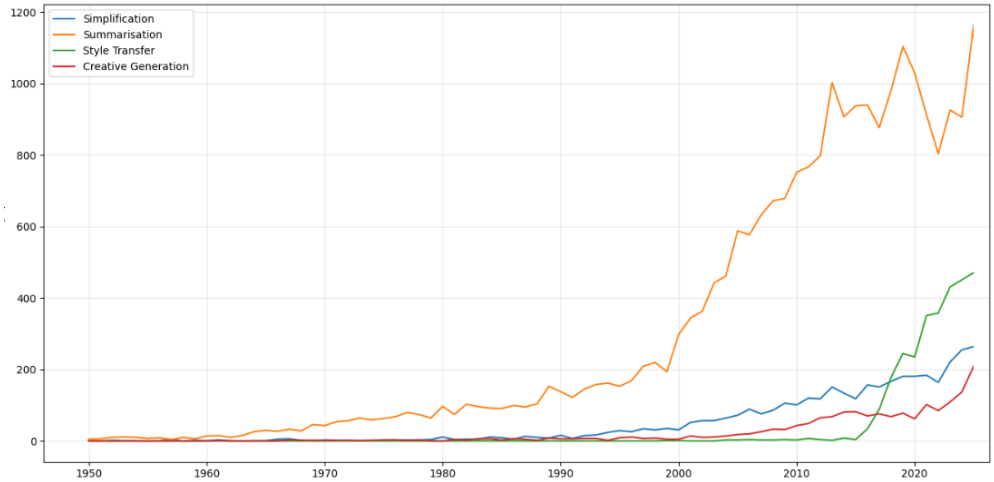

Figure 2
Interest in the four investigated NLP tasks, represented by the number of associated articles per year in Semantic Scholar.

In particular, simplification and creative generation demonstrate very similar patterns. A recent phenomenon, style transfer emerged around this same time and as of today has surpassed both simplification and creative generation in terms of article presence.

We will commence the analysis by presenting the need for automatic literary adaptation and introducing its traditional, human-led counterpart. Then, in sections 4-7, we will focus on each of the four mentioned related tasks (simplification, summarisation, style transfer, and creative generation). Each of these sections will comprise: 1) a definition of the task, 2) a general timeline of its development (typically moving from rule-based methods through statistical and neural methods to reach LLMs as the current state of the art), 3) associated evaluation methods and 4) takeaways in view of automatic literary adaptation. Our focal points will include evaluation practices, multilingualism and the use of LLMs. We will end with a discussion that sums up the practices, strengths and limitations of the analysed NLP subfields and sets general guidelines concerning the steps that should be taken in order for automatic literary adaptation to be defined as a task in its own right. Throughout the study, we will use the term "literary adaptation" rather than "literary simplification" in order to underline the complex and multifaceted nature of altering text to address a large variety of readerships.

2. Motivation

The reading of literature has been shown to aid the development of key psychological and social abilities ranging from critical thinking and imagination to emotional understanding and empathy (Nussbaum 2010; Oatley 2011; Mar et al. 2006). With their experimental study

conducted with over 33,000 participants, Mar et al. (2021) found that narrative text is better understood and memorised than essay-like text. However, there is also evidence that when a reader does not sufficiently understand a literary text, the benefits are at best reduced and, at worst, there may be harmful consequences for the reader, such as a decrease in confidence and motivation (Krashen 1982; Nation 2001; Day and Bamford 1998). In contrast, the matching of text to one's current abilities has been shown to result in better learning outcomes (FitzGerald et al. 2018). The achieved higher reader interest reduces "mind wandering", leading to higher engagement and even understanding across ages, languages, and text types (Bonifacci et al. 2023). Accordingly, the past few decades have seen significant production of specially conceived literary works, which in turn more or less narrowly address the needs of defined audiences. These works are often adaptations of already existing and popular (a.k.a. "canonical") texts. Specific book series and editions have been developed that are for children, learners of foreign languages, and people with learning disabilities. The most established ones follow clear frameworks that separate their audience in a fine-grained manner in accordance with their needs, such by specific ages (e.g. Oxford Reading Tree, Penguin Young Readers) or specific proficiency levels (e.g. Oxford Bookworms for English, Lectures CLE en français facile for French).

Still, the availability of adapted literature is limited by a series of factors, including the predominance of English and other popular languages at the expense of most others, concerns of physical availability and affordability, the multiplicity of differentiated audiences and even the implied workload required from professionals in order to compose the texts. Given the current technological context, it is therefore natural to envision the assistance of automated tools, whose extent remains to be determined based on technological potential and ethical norms. For the purpose of gently addressing both of these concerns, in the current paper we firstly raise the question of adaptation of already existing literary texts rather than the composition of new ones.

3. Background: Literary Adaptation

Hutcheon (2006) points out that the word "adaptation" denotes both a result or artefact and the process of its creation. Ewers (2009, p. 179) goes on to say that the practice signals "any modification of a work as it moves across boundaries of media or genres". Occasionally, more than one such movement is accomplished at a time; for instance, in its versions for children, the novel *Gulliver's Travels* typically shifts from satire to adventure story. Other elements that may change in the process of adaptation include points of view, ontologies and philosophies (Hutcheon 2006). Artistic adaptation has a long history that predates modernity; for instance, in Victorian times, stories, poems, operas, paintings and other media would often allude to one another.

Specifically addressing literary adaptation, Lefebvre (2021, p. 54) defines it as "texte rémanent d'un texte littéraire source, socialement reconnu dont il a pour fonction de tenir lieu" ("a text deriving from a socially recognised source literary text, whose function is to stand in its place"[3]). The extent of the undergone changes may vary significantly in size

3 Translations are the first author's.

and nature. Nières-Chevrel (2009, p. 190) ironically notes that children's literature "se donne toute liberté de ramener à vingt pages un roman qui en faisait deux cents, de maintenir ou de taire le nom de l'auteur de l'œuvre originale pour n'en garder que le seul titre" ("allows itself all freedom to bring down a two-hundred-page novel to twenty pages, to retain the name of the original author or to omit it, keeping only the title"). Occasionally, it may be the original author that adapts their own text, such as in the case of Lewis Carroll's *Alices' Adventures in Wonderland* and its two later versions, including one markedly meant for children (entitled *The Nursery Alice*). One may notice that it is often recurring canonical literary works (e.g. *Pinocchio*, *Don Quixote*, *Les Miserables*) that are subject to adaptations into films, children's stories and other media and genres. Ellis (1982, p. 190) points out that such works have been so recycled through formal adaptation and other allusions that the audience is now familiar with "a generally circulated cultural memory" rather than the original texts. Considerations in addition to (but often related to) canonical status that may influence the choice of literary works to undergo adaptation include financial planning and the demands of educational systems.

The canonical texts that are subject to literary adaptation often carry cultural significance for their target audience e.g. young children that share the text's underlying culture or language learners that may be interested in discovering not only the target language but its associated culture. The status of cultural heritage may cause adverse reactions to the process of adaptation and equate it to a loss in authenticity. An example of the phenomenon occurred in relation to author Russi Chanev's adaptation of the famous nineteenth-century Bulgarian novel *Under the Yoke*, which highly altered the original archaic language of the work in order to increase its comprehensibility for children of Bulgarian origins who live outside of the country (Vazov 2024; Intrigi.bg 2024). Addressing potential issues of the type, Hutcheon (2006, p. 70) claims that "[t]here always is a trade-off in adaptation". Thiel (2013) qualifies said trade-off in the following way: whilst an adaptation might decrease an original text's historical value, it offers contemplation on the relationship between the past and the present in relation to literature and beyond. Postmodern and poststructuralist views of adaptation include a shift away from its secondary status and towards its importance as an independent artefact (Müller 2013).

We will now discuss several key audiences that tend to be addressed in the process of literary adaptation. Children constitute the most largest one. Adapted texts for children typically have higher availability compared to those for other audiences, notably when it comes to low-resourced languages. The practice of adaptation for children has a century-long history. Müller (2013) cites Joachim Heinrich Campe's *Robinson the Younger* of 1780 (an adaptation of *Robinson Crusoe*) as the first formal literary adaptation for a young audience. Other early adaptions include Thomas Bowdler's *Family Shakespeare* (1807), a children-friendly version of Shakespeare's plays. The need for adaptations for children has historically stemmed from a general focus on education as well as on the idea that children need to be protected from explicit or traumatic material. In the current landscape, the link between literature and education is sometimes rethought, and there is a vast variety of books to choose from in accordance with a child's specific interests (Shavit 1999). In addition to being adapted, children's classics often reach their audience as translations, thereby introducing

another layer of "adaptation" (one to a different language and culture). Also, an older source text may be seen as simultaneously adapted to another time period i.e. modernised.

There are existing attempts to methodologically define and even quantify the practice of literary adaptation for children. Van Lierop-Debrauwer (2000) categorise the adjustable aspects of a literary work as content, style, structure and design. Providing a microscopic view on ideational grammatical metaphors (the abstract in nature expression as a noun of a notion that would typically require a verb or adjective) in literary adaptation for a younger audience, Liu (2021) defines the following five strategies: addition, maintenance, revision, unpacking and demetaphorisation. In a curious study that compares texts written in English and Dutch for children, teenagers and adults by the same authors (referred to as "crosswriters"), Haverals, Geybels, and Joosen (2022) apply stylometric features, clustering and PCA analysis to ultimately conclude that literature for children shares more similarities than works for different audiences by the same author. This conclusion suggests that audience-driven variation is a measurable and systematic phenomenon.

In the past few decades, practices such as comprehensive input and extensive reading have been dominant in foreign language education (Krashen 1982). They are directly related to observations such as the easier acquisition of vocabulary items when presented within natural context (Godwin-Jones 2018; Restrepo Ramos 2015). Literary series, known as "graded readers", have emerged in an attempt to maximise the learning process. Used in the context of both independent and in-class learning, these materials have proved to help reaffirm vocabulary knowledge as well as increase student motivation and sense of community (Wan-a rom 2008; Hill 2013; Dickinson 2017). Graded readers often consist in adaptations of famous literary texts, typically related to the target language's culture. Their difficulty is classified in terms of proficiency levels, such as per the Common European Framework of Reference for Languages (CEFR). The expected number of "headwords" i.e. already acquired vocabulary knowledge, may also be indicated. The literary text may be accompanied by glossaries, exercises, and cultural information. Today, graded reader series in English, such as Oxford Bookworms, Penguin Readers, and Cambridge English Readers are an established phenomenon. Albeit to a much lesser extent and proportionately to their general popularity/resourcedness, other languages also benefit from such materials; for instance, Lectures CLE en français facile (French), Leichtzulesen (German), Sistema Kalinka (Russian), Mandarin Companion (Chinese), and Taishukan (Japanese).

Specific learning disabilities and other medical conditions also influence a person's understanding of written text (and literary text in particular) in different ways, which are more or less narrowly related to the condition at hand. Autistic people tend to have difficulty in processing figurative language (Happé 1993; Kalandadze et al. 2018) that may be improved given relevant intervention such as the inclusion of images or explicit explanations (Melogno, Pinto, and Orsolini 2017; Rutherford et al. 2020; Mashal and Kasirer 2012). Dyslexia is also often associated with reduced understanding of abstract language as well as long words and complex syntax (Snowling, Hulme, and Nation 2000; Cersosimo, Domaneschi, and Cancer 2025). While not intrinsically related to language acquisition and understanding, ADHD similarly impedes the processing of long segments and complex syntax, and patients benefit from the inclusion of explicit signalling, such as headings and keywords (Martinussen et al. 2005). In turn, aphasia has been associated with difficulties

in relation to infrequent words and ambiguous syntactic structures (Caplan 2012; Dickey and Thompson 2009). While to our knowledge, there are no established literary editions or series that target specific disabilities, it is worth mentioning examples such as the UK-based publisher Every Cherry, which defines its audience via the wider term "special needs" as well as Spanish publisher Almadraba's collection Kalafate, which implements practices of easy-read, thereby addressing conditions such as dyslexia and ADHD.

4. Related NLP Tasks: Simplification

4.1 Definition

The purpose of automatic simplification is to provide a simpler version of a text that retains efficiently its original information. The output should also be coherent and grammatical. The new text may be addressed at a reader audience or used as a pre-processing step in the context of additional NLP tasks. A common division is that between lexical and syntactic simplification. Utilised techniques may include syntactic operations (e.g. sentence splitting), lexical operations (substitution of infrequent or technical words), content reduction (through summarisation and paraphrases) and clarification (added explanations or definitions) (Štajner and Saggion 2013). Recent large-scale initiatives aiming at text simplification (such as the US Government's Plain Language Action and Information Network[4] and Inclusion Europe's Easy-to-Read Guidelines[5]) have placed a focus and a sense of urgency on the need for efficient automation of the process.

4.2 Timeline

Rule-based text simplification typically includes the application of manually-established syntax rules and the replacement of vocabulary items based on pre-established lists of synonyms. Devlin and Tait composed parse trees of sentences, based on which rules of splitting, rearrangement and deletion were applied. Devlin (1999) and Devlin and Unthank (2006) specifically developed rules targeted at the increase of text understandability by aphasia patients.

The launch of Simple Wikipedia in 2003 became a turning point in statistical text simplification, as it provided numerous ready examples of paired original and simplified text. Using aligned sentences from the original and simple Wikipedia, Coster and Kauchak (2011a) defined the simplification task as an English-to-English translation task. The translation method was to also be applied to additional languages and dialects, such as German, Chinese, Swedish and Basque (Klaper, Ebling, and Volk 2013; Chen et al. 2012; Stymne et al. 2013; Aranzabe, Ilarraza, and Gonzalez-Dios 2012). Biran, Brody, and Elhadad (2011)'s unsupervised model learned simplification rules from a comparable corpus. In contrast, Glavaš and Štajner (2015) used word embeddings and cosine similarity in the task of lexical

4 https://www.plainlanguage.gov
5 https://www.inclusion-europe.eu/easy-to-read/

simplification in addition to WordNet lists, frequency and a classifier model that seeks words with highest simplicity. Coming first at the relevant SemEval 2016 Complex Word Identification task, Paetzold and Specia (2016) combined different lexicon-based, threshold-based and machine learning approaches. Stajner and Saggion (2013)'s work on automatic text simplification was focused on the Spanish language and addressed discrete reader audiences, such as people with Down's syndrome and autism spectrum disorders. They used a corpus of professionally adapted articles on different topics and applied the operations of sentence deletion and splitting.

The introduction of neural NLP offered new state-of-the-art simplification techniques. Nisioi et al. (2017)'s sequence-to-sequence simplification model achieved higher grammaticality and meaning preservation as well as a higher degree of simplification compared to earlier methods. Zhang and Lapata (2017)'s reinforcement-learning-based DRESS (Deep REinforcement Sentence Simplification) model was trained to encourage output that preserves the meaning of original input while maintaining simplicity and fluency. Cripwell, Legrand, and Gardent (2023) proposed a high-performing transformer model that is based on document-level simplification (in contrast to the common sentence-level). The model marked sentences to be copied, rephrased, split or deleted based on both context and internal structure.

The LLM era has been marked with a variety of projects linked to text simplification. Kew et al. (2023) proposed Benchmarking Large Language Models on Sentence Simplification (BLESS), a comparison of 44 LLMs ranging from 60 million to 176 billion parameters in relation to the task. They used a few-shot setting and applied both qualitative and automatic analysis (including SARI, BERTScore and LENS). Closed-weight models, in particular Davinci-003 and GPT-3.5-Turbo, performed particularly well, and instruction-tuning was shown to improve models' performance. Feng et al. (2023) concluded that GPT3.5 and ChatGPT perform the task as well as humans in English, Portuguese and Spanish. Evaluation was based on SARI and FKGL[6] (and the latter's Spanish counterpart, FRES); in addition, 100 sentences were human-evaluated. Yang (2024) compared fine-tuning and few-shot prompting of GPT models on the simplification of medical-related sentences using a professional dataset and human evaluation of correctness and quality, reaching the conclusion that the two methods were comparable in performance. Ormaechea and Tsourakis (2024) fine-tuned FLANT5-based language models for the simplification of French text, using the WiViCo dataset (which contains over 40k pairs of complex and simpler sentences) and evaluated them using human evaluation, BLEU and SARI scores. The BEA 2024 shared task approached multilingual lexical simplification as covering a variety of textual genres and target audiences (Shardlow et al. 2024). The highest-performing models used a variety of techniques ranging from handcrafted rules to LLM models such as GPT, proving the former's ongoing relevance.

Recent advancements in automatic text simplification have also led to finer-grained reader audience distinctions, such as simplification to a particular CEFR level. Farajidizaji, Raina, and Gales (2024) tested the ability of the popular models ChatGPT and Llama-2

6 Flesch–Kincaid Grade Level, a readability measure, see (DuBay 2004)

to bring text of any source level to any target level in terms of Flesch Reading Ease[7]. Indeed, Spearman correlation proved that the models succeeded in adjusting the readability of documents. However, the issuing levels did not map unambiguously to Flesch values; instead, they were dependent on the source texts' readability. Similarly, Benedetto and Buttery (2025) noted that a selection of LLMs, including Llama-3 and GPT, perform CEFR-targeted simplification efficiently yet imperfectly, typically oversimplifying texts. Jamet, Shrestha, and Vlachos (2024) approached French-language simplification based on CEFR levels, focusing on current models' ability to reduce a source sentence by a single level at a time. They used labelled sentences in a few-shot setting and classified the derived text's level using the CamemBERT model. GPT-4 achieved best results as compared to Mistral and Davinci models.

4.3 Evaluation

Alva-Manchego et al. (2020) sum up that the following main criteria are considered at evaluation of automatic textual simplification models: output fluency, meaning preservation and simplicity in comparison to the input (unsimplified) text. Many authors opt for automatic scores conceived for the task at hand or for related NLG tasks.

Translation Edit Rate-plus (TERp) provides the number of edits required in order to transform the source text into the simplified output (Snover et al. 2009). Designed specifically for sentence simplification, SARI (Xu et al. 2016) measures accuracy and recall in relation to words that have been added, deleted or retained, making use of several reference simplified sentences. Originally used for machine translation, the term-overlap metric BLEU measures n-gram overlap between the reference and evaluated texts, while penalising output deemed too short (Papineni et al. 2002). Adjustements of BLEU introduced in view of the task of simplification include FKBLEU (Xu et al. 2016), which compares the output to both a reference and the input. BLEU has been noted to perform better in terms of meaning preservation, whilst SARI is a better estimate for simplicity; however, both metrics show overall low correlation with human judgement (Xu et al. 2016; Al-Thanyyan and Azmi 2021). Also used primarily in machine translation tasks, METEOR improves on BLEU by calculating matches to the reference based on unigrams' surface forms, stemmed forms and meanings, while also taking into consideration the order of matched elements (Banerjee and Lavie 2005). Another metric derived from BLEU is NIST, which assigns different weights to n-grams based on their informativeness (Doddington 2002). Originally used for machine translation and widely applied to text simplification, COMET, state of the art on the WMT 2019 Metrics shared task, has an XLM-RoBERTa encoder and makes use of the source text and a multilingual embedding space (Rei et al. 2020). ROUGE (Lin 2004) is another relevant n-gram-based metric, which was conceived for the task of summarisation (see Section 5.3 for more detail). It has, however, been noted to correlate poorly to human judgement at the task of simplification (Scialom et al. 2021a). Demonstrating better correlation with human judgement, SAMSA may be seen as an improvement on SARI that provides a better overview

7 See DuBay (2004)

of text simplification results by addressing semantic structure (Sulem, Abend, and Rappoport 2018). It is a reference-less metric that focuses on predicate-argument relations while considering "scenes" rather than sentences. BERTScore changes the evaluation landscape by making use of a BERT model to calculate semantic rather than n-gram-based similarity between an automatic output and a reference text (Zhang et al. 2020). It is applied to a variety of NLG tasks, including simplification.

QuestEval, formulated for summarisation tasks, was later updated so as to be relevant to simplification evaluation (Scialom et al. 2021b). The metric evaluates the factual consistency of an output based on the associated source document. For the purpose, a list of questions is generated based on the former, and answers are then retrieved from both texts, whilst their similarity is calculated. In its simplification-friendly version, the similarity is calculated using BERTScore instead of the original F1-score, thereby accounting for the use of synonyms and reformulations. Finally, LENS is a metric specifically crafted for simplification that uses a pre-trained model, based on human judgements pertaining to meaning preservation, simplicity and fluency (Maddela et al. 2022). The EASSE framework (Alva-Manchego et al. 2019) standardises the evaluation of text simplification through commonly used automatic metrics, offering a unified toolkit. In turn, Stodden (2024) propose EASSE-multi, a related framework that extends to non-English languages, and exemplify its use for German simplification.

Many of the mentioned automatic metrics used in text simplification are dependent on the presence of high-quality datasets of associated simple and complex sentences. Popular datasets of the type include Wikipedia-Simple Wikipedia for English (Coster and Kauchak 2011b), Newsela for English and Spanish (Xu, Callison-Burch, and Napoles 2015), PorSimples for Brazilian Portuguese (Aluísio and Gasperin 2010), Alector for French (Gala et al. 2020) and 20 Minuten for German (Rios et al. 2021). OneStopEnglish is a corpus for readability assessment and automatic text simplification that consists of 189 original texts, each accompanied with three graded simplifications for learners of English as a second language(Vajjala and Lučić 2018). It is worth mentioning that datasets focused on literary simplification, although limited in number and size, have also been produced. Washburne and Vogel (1926) published the *Winnetka Graded Book List*, which includes 700 children's books, manually annotated for reading difficulty. MULTISIM is a multilingual simplification benchmark that includes 1.7 million pairs of complex and simplified sentences over 12 languages(Ryan, Naous, and Xu 2023). It covers eight domains, including literature (novels). Brunato et al. (2015) offer a corpus of 32 short Italian novels for children and their manually simplified versions for "poor comprehenders''. They go on to annotate it for a selection of simplification procedures[8]. Dmitrieva and Tiedemann (2021) compile a simplification dataset of 19 books for Russian-language learners at different CEFR levels; then, they go on to train an LSTM-based classification model and evaluate it using the EASSE framework.

Constituting a differing approach, evaluation through measures of readability is often evoked in relation to automatic simplification. The term "readability", in use since the 1950s, denotes measures and formulas that are meant to estimate a text's level of difficulty, typically

8 split, merge, reordering, insert, delete and transformation

so as to link it to a particular reader audience, such as a school grade level (DuBay 2004). Early readability formulas focus on length-based measures such as word and sentence length (e.g. Flesch Reading Ease) and/or on lexical mapping (e.g. Dale-Chall). Originally conceived with the English language in mind, readability formulas based on superficial text characteristics have been shown to correlate across languages (van Oosten, Tanghe, and Hoste 2010). Also, adaptations of some readability formulas for other languages have been elaborated, such as Flesch-Douma (an adaptation of Flesch Reading Ease for Dutch) (Douma 1960) and Spaulding's Spanish readability formula (an adaptation of Dale-Chall) (Spaulding 1956). The main advantage of readability-based features and formulas is that they make up an universal convention that is not dependent on a limited reference corpus. They can therefore be applied independently in evaluating the difficulty of both an original text and a simplified text derived from it. However, the method is also associated with significant limitations, such as the inability to account for errors and to address textual semantics. Less "shallow" atomic textual features not typically present in readability formulas have also been noted to have high relevance to textual simplification, such as the use of complex sentences, passive and conditional constructions, and figurative expressions (Štajner and Saggion 2013; Evans, Orăsan, and Dornescu 2014).

Human-based evaluation has also been widely applied in automatic simplification tasks, with a common focus on fluency, grammaticality, simplicity, and meaning preservation. Al-Thanyyan and Azmi (2021) use a Likert scale (1 to 5 or 1 to 3) to estimate the mentioned features. Maddela et al. (2022) introduce the human evaluation framework RANK & RATE, in which candidate simplifications are rated through an interactive interface.

4.4 Takeaways

Automatic text simplification is possibly the discussed task with highest relevance to automatic literary adaptation, and many of its developments and conclusions are readily applicable to the latter. Initially, the task tended to involve no clear definition of "simple" and "complex" text, and the focus was on the difference in simplicity between source and target text rather than a more universally applicable distinction. Gradually, narrow reader audiences such as people with different disabilities and foreign language learners at different levels came to be more frequently addressed in simplification research. Recent advancements, such as document-level granularity and application to various languages, although still developing, are cause for optimism. They go hand-in-hand with general NLP trends, such as language models' increasing semantic knowledge and proficiency in low-resource languages.

The evaluation of simplified text via readability-based and other predefined characteristics has the advantage of alleviating the need for a large number of reference datasets associated with different textual genres, audiences or languages. Established automatic measures, while imperfect individually, have the potential of covering different textual aspects (e.g. SARI has proven to be a good proxy for lexical simplicity) and to therefore work efficiently when used collectively. Finally, human evaluation has been used consistently with text simplification, including in the era of LLMs and is likely to continue being a necessity.

Future advances, including in literary adaptation, can benefit from a systematisation of efficient human evaluation frameworks.

5. Related NLP Tasks: Summarisation

5.1 Definition

As Fig. 2 demonstrates, interest in automatic summarisation has been both early and widespread. The need for reduction in textual size goes hand in hand with the increase in quantity of online data and people's related difficulty in consuming it in terms of time and effort. Summarisation is often an element of simplification, although the two tasks can also exist independently of one another (as shown, the simplification of a text may sometimes require its increase in textual length, such as through the addition of explanations). Although the decrease in size involved in summarisation may vary significantly based on the associated documents and tasks at hand, a general guideline is that reduction by at least half is to be aimed at (Radev, Hovy, and McKeown 2002). The most widely accepted categorisation is that between the extractive and abstractive approaches to summarisation. In the case of the former, the parts of a text that are of highest importance are taken verbatim to constitute its summarised version. In contrast, the latter approach implies reformulation of the text. Earlier practices of automatic summarisation typically followed the extractive approach and involved preprocessing the source text and ranking its tokens (usually, sentences) in order of importance, such as through TF-IDF-based methods. In turn, as abstractive summarisation involves novel text generation, it became widespread during the pre-neural and neural periods. Still, this is not to say that the need for extractive summarisation disappeared, nor that abstractive summarisation was not attempted earlier on. A distinction also exists between generic and query-based summarisation. Generic summarisation seeks to provide a general overview of a text's content, whilst a query-based one focuses on information that is relevant to a specific search query at hand. In addition, summarisation may be of one or multiple documents. Like simplification, summarisation may also be used as an intermediary step in other NLP tasks, such as information retrieval and question answering.

5.2 Timeline

Luhn (1958)'s work on automatic generation of articles' abstracts has been cited as the first application of automatic text summarisation. He made use of word frequency and distribution in order to rank the significance of both words and sentences. Hu and Liu (2004) provided summaries of the totality of customer reviews of a given product based on the mentioned features of the product as well as a review's overall positivity or negativity. Ko and Seo (2008) combined pairs of consecutive sentences prior to ranking them for significance so as to account for context. Addressing a literary audience, Kazantseva and Szpakowicz (2010) extracted summaries of short stories with the purpose of aiding readers in deciding whether to read the full text. Heuristics were used to select descriptive content that pertains to a story's setting and main entities but does not reveal the plot. Genest and

Lapalme (2012) combined rule-based methods of information extraction with NLG in the production of summaries.

During the pre-neural NLP period, extractive summarisation methods still dominated the field. Ouyang et al. (2011) used Support Vector Regression (SVR) to estimate the importance of sentences in the context of query-focused multi-document summarisation. Wang and Ma (2013) represented texts as matrices and inferred underlying topics via dimensionality reduction, achieving high ROUGE-score results. In turn, Shetty and Kallimani (2017) applied matrix decomposition and k-means clustering of sentences in order to extract diverse and redundancy-free summaries. Kobayashi, Noguchi, and Yatsuka (2015) used word embeddings to represent documents and candidate summaries and ranked different extractive summaries based on distances between embeddings.

Due to its simplicity and efficiency, extractive summarisation did not become obsolete even with the rise of neural NLP. Cheng and Lapata (2016) approached the task using a hierarchical encoder that models document structure combined with an attention mechanism that determines tokens' importance for a given document. The model was trained on a large dataset and did not imply textual preprocessing. Abdel-Salam and Rafea (2022) evaluated BERT models on the task of extractive summarisation and proposed the fine-tuned easily trainable "SqueezeBERTSum" model.

Attention-based sequence-to-sequence models proved to be particularly fitting for the task of automatic summarisation, particulary in relation to long documents. Chopra, Auli, and Rush (2016) trained a RNN model with an attention-based encoder to aid the decoder's focus at the generation of abstractive summaries. Chen et al. (2019) proposed RC-Transformer (RCT), a model that works efficiently with long-term dependencies and outputs better-ordered text. The model was extended with an additional RNN-based encoder and a convolution module that filters text with local importance.

The LLM era brought about a variety of summarisation-related work and an exploration of methods such as prompt engineering, fine-tuning and knowledge distillation (Zhang et al. 2026). The primary focus remained on informativeness and the reduction of large amounts of data for practical purposes. Xiao and Chen (2023) proposed an LLM-based method of evolutionary fine-tuning of news articles, focusing on informativeness. Human evaluation ranked LLM output as better than one produced by humans and fine-tuned neural models, specifically emphasising its qualities of fluency and coherence. Zhang et al. (2024) compared performance at the task between humans and ten discrete LLM models, concluding that the former's summaries are more abstractive in nature. Pu, Gao, and Wan (2023) boldly claimed in their article's very title that (human-based) "summarization is (almost) dead", basing themselves on the results of five summarisation tasks, including crosslingual summarisation. In contrast, Wang et al. (2023a) had significant criticism regarding current LLMs' crosslingual summarisation abilities. More specifically, they noted that GPT-4 demonstrates competitive performance based on ROUGE and BERTScore (but still performs worse compared to a fine-tuned BART model), whilst Vicuna-13B outright lacks zero-shot abilities for the task. They pointed out that the use of chain-of-thought, such as instructions to first translate and then summarise a document, helps improve models' performance.

The potential of current models to summarise long text is particularly relevant in view of literary adaptation. Wu et al. (2024) incrementally extracted and concatenated key

sentences from a long document, stopping the process when the derived summary resulted in the highest ROUGE score. Chang et al. (2023) divided a long document of over 100k tokens into chunks, and the chunks were then merged hierarchically and incrementally, while textual coherence was used to evaluate the resulting summaries. The highest scoring models were GPT-4 and Claude 2, and hierarchical merging achieved better results.

5.3 Evaluation

The main qualities brought forward in relation to automatically summarised text include its information coverage, information significance (general or user-/task-defined), coherence and lack of redundancy (Huang et al. 2010). Human-based evaluation has focused on these criteria, typically as organised in ranking scales (El-Kassas et al. 2021). Many of the automatic metrics used for the task, including term-overlap based ROUGE, BLEU and METEOR and semantic-similarity-based BERTScore, were already mentioned in relation to automatic simplification (see Section 4.3).

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric was conceived for the task of summarisation (Lin 2004). In its framework, an automatic summary is compared to one or more references in terms of overlapping units. Variations of ROUGE include ROUGE-N, which is based on n-gram units; ROUGE-L, which denotes the longest common subsequence; and ROUGE-S, which offers skip-bigram statistics, thus allowing for gaps between tokens. The ROUGE metric focuses on recall, whilst BLEU (typically used in the evaluation of machine translation), focuses on precision. Metrics that have been introduced in order to improve on ROUGE's performance include Basic Elements, where sentences are broken down into small meaning units, against which summaries are compared (Hovy et al. 2006) and BEwT-E (extended Basic Elements), which also accounts for textual transformations, such as paraphrases and syntactic variations (Tratz and Hovy 2008). In turn, DEPEVAL makes use of dependency parsing and compares summaries based on syntactic structures (Owczarzak 2009).

Metrics such as BERTScore, which take into account textual semantics, are often seen as a better evaluation alternative of summarisation output, especially in cases of abstractive summarisation. Yet, BERTScore has been noted not to offer significant benefits when compared to ROUGE (Scialom et al. 2021a). BLEURT is another similarity score based on BERT, which models human judgement. It can be used in cases of limited reference data; however, it requires training on existing human ratings (Sellam, Das, and Parikh 2020). BARTScore is a comprehensive framework that makes use of a pre-trained BART sequence-to-sequence model (Yuan, Neubig, and Liu 2021). Its main idea is that higher-quality summaries receive better scores based on how likely they are to be generated from, or to generate, a reference text. The score has been noted to correlate well with fluency and accuracy. The Pyramid method is a human-centric evaluation framework that weighs Summary Content Units (units that represent key information) across multiple reference summaries (Nenkova, Passonneau, and McKeown 2007).

Scialom et al. (2021a) proposed QuestEval, a question-answer-based method that does not require reference summaries or ratings at evaluating whether a summary contains all relevant information from an associated source document. QuestEval correlates highly with

human judgement in terms of consistency, coherence, fluency, and relevance. Recently, LLMs have also been made use of as evaluators of automatic summaries. Jain et al. (2023) explored ChatGPT's ability to replicate evaluation methods such as Pyramid and pairwise comparison, achieving high results. Chang et al. (2023)'s BOOOOKSCORE uses GPT-4 prompts to identify problems within automatic output, including causal omissions, salience, and duplication.

Large reference datasets used in the evaluation of automatic summarisation include Gigaword (Napoles, Gormley, and Van Durme 2012) and DUC (National Institute of Standards and Technology 2004)[9], LCSTS (Hu, Chen, and Zhu 2015)[10] and Wikilingua (Ladhak et al. 2020)[11].

5.4 Takeaways

Literary adaptation typically involves an element of summarisation i.e. a significant reduction in size. However, a key distinguishing feature in literary adaptation is that unlike in the case of most examples of automatic summarisation, the focus therein lies in the reading experience itself rather than the efficient presentation and consumption of information. As a result, additional questions arise that do not tend to be addressed within the generation and evaluation of automatic summaries, such as those of general aesthetics, narrative voice, and anaphora resolution.

Abstractive summarisation is the method that holds greater relevance in relation to the literary domain and, as with simplification, recent advancements, such as increasing efficient work with long context and a focus on semantics, speak of possible "readiness" for NLP to engage in the task. Current evaluation methods for automatic summarisation that may be applicable to literary adaptation include those that involve a multitude of features and rely on a common semantic space, thereby being largely language-independent.

6. Related NLP Tasks: Style Transfer

6.1 Definition

The term "style transfer" or "style adaptation" may be used in relation to different modalities, such as image, audio and textual data. For the purpose of this survey, we will be addressing solely textual style transfer i.e. an NLP task that consists in altering the style of a text whilst preserving its semantic content. The definition of textual style is largely open to interpretation, and the task may so much as be taken as an umbrella term that includes the aforementioned tasks of simplification and summarisation. Tikhonov and Yamshchikov (2018) note that style needs to be "orthogonal" in relation to semantics in the sense that any semantic quality of a text can be expressed in any defined style. While admitting subjectivity,

9 news summaries in English
10 blog articles in Chinese
11 a multilingual knowledge base

Xu (2017) refers to the following seven key styles: "simple and short", "instructional and robotic", "historical and evolving", "colloquial and internet", "gendered and personalised", "pervasive and framing" and "polite and abusive". The most commonly approached dichotomies within traditional style transfer tasks are those between formal and informal text and positive and negative sentiment. Additional focus may fall on authorial voice, register and audience adaptation. Applications of style transfer include the generation of personalised texts and consistent dialogue systems. Style transfer may also be used within other NLP tasks, such as data augmentation. In their extensive survey on the topic of textual style transfer, Jin et al. (2022) note that the rising interest in style within NLP is of high significance, as it implies unprecedented user-centredness.

6.2 Timeline

As noticeable in Fig.2, the practice and interest in style transfer were very scarce during the pre-neural period. The limited number of rule-based studies typically involved the application of hand-crafted grammars and templates. Reddy and Knight (2016) experimented with the alternation of authorial gender through lexical substitution for the purpose of concealing identity in social media communication. Working with the Japanese language, Mizukami et al. (2015, p. 129) defined style transfer as a statistical machine translation task that seeks to "express the individuality of the writer or speaker".

The emergence of neural NLP opened new potential in relation to the examined task. Inspired by recent advances in style transfer within computer vision, Zhang, Zhao, and LeCun (2015) used ConvNets (temporal convolution networks) to deconstruct English and Chinese text at different granularities ranging from character-level to abstract textual features. Addressing sentiment transfer, Li et al. (2018) were among the first to seek to disentangle stylistic from content-related textual attributes in an unsupervised way. They then deleted and added from the original text phrases determined to be distinctive to positive or negative sentiment. Aiming to increase the politeness of chatbot replies, Mukherjee, Hudeček, and Dušek (2023) addressed the unavailability of training data by generating synthetic data via the intermediary of a politeness transfer model.

An example of the widespread application of adversarial networks in style transfer is the work of Zhao et al. (2018). They made use of an adversarially regularized autoencoder (ARAE) that jointly trains a rich discrete-space encoder (e.g. an RNN) and a continuous space generator function, constraining the resulting distributions to be similar. In turn, Fu et al. (2018) defined style as a set of measurable categorial and continuous parameters and used adversarial networks to learn separate content and style representations. They evaluated output based on transfer strength (the extent to which the target style is reflected) and content preservation (the extent to which the original content is retained).

Key training data assembled and/or utilised efficiently within the task of style transfer includes Jhamtani et al. (2017)'s parallel corpus of Shakespearean versus modern language, used to train a "translator" between the two language styles. They used a bidirectional LSTM as an encoder and a mixed RNN and pointer network module that enables copy action as a decoder. Simultaneously undertaking the tasks of automatic translation and style transfer, García et al. (2021) defined both sentiment and language as textual attributes that may be

modified jointly. Surya et al. (2019) also tackled several NLP tasks at applying style-transfer techniques for the purpose of content reduction and lexical simplification. They used an adversarial setup and several auxiliary losses. Addressing this landscape of numerous and varying neural style transfer methods and projects, Tikhonov and Yamshchikov (2018) published an opinion paper that served as a quest for formalisation and standardisation of the task in terms of definition, approaches and evaluation.

Naturally, LLMs brought an array of new approaches to the practice of textual style adaptation. Reif et al. (2022, p. 837) focused on models' zero-shot abilities at a sentence rewriting task, achieving good results in relation to sentiment transfer as well as "transformations such as 'make this melodramatic' or 'insert a metaphor.' " In particular, ACL's 2024 edition was centred around controllable generation and introduced a variety of works that tackle the examined task. Liu et al. (2024) addressed the issues of errors within output and lack of user control in their prompt-based approach for specific-region editing. LLMs were instructed to modify limited portions of text within a defined editing region. Region selection was based on prompt engineering as well as frequency-based strategies. Employing LLMs and chain-of-thought prompting, Han et al. (2024) generated synthetic parallel data for the task of style transfer, on which they then trained a model that performs style-content disentanglement, introducing two losses in order to focus on attribute-related features while constraining a text's semantic space. In their recent work, Huynh and McNamara (2025) interestingly evoked the term "textual personalisation" rather than "style transfer", thus offering an emphasis on the reader's characteristics. They prompted several LLMs to adapt ten scientific texts in accordance with several reader profiles that were defined in terms of age, educational background, reading skills, interests and learning goals. Prompts were augmented using chain-of-thought, personification and RAG techniques. The authors noted that evaluation of the achieved modifications was a rather challenging task. Through automatic metrics, Llama was concluded to produce the most complex and academic-like texts, Gemini - the most diverse ones in terms of both vocabulary and syntax, whilst Claude - the least cohesive ones.

6.3 Evaluation

Evaluation of style transfer output is challenging due to the markedly non-quantitative nature of "style". Jin et al. (2022) suggest that, despite the current technologically advanced landscape, the task necessitates both human-based and automatic evaluation.

The following qualities are typically evoked in relation to style transfer evaluation: fluency, transfer strength, and content preservation (Fu et al. 2018). The last measure has typically been evaluated with established n-gram-based and semantic-based similarity metrics, such as BLEU, ROUGE, METEOR, and BERTScore. However, there is a general lack of associated reference datasets for the task. Popular parallel style-transfer datasets include GYAFC (Grammarly Yahoo Answers Formality Corpus) for formality transfer (Rao and Tetreault 2018), the Yelp Reviews dataset for sentiment transfer (Shen et al. 2017), the Shakespeare dataset of modern versus Shakespearean style (Xu et al. 2012), and CDS (Corpus of Diverse Styles), which involves a plurality of defined styles (e.g. poetry, Biblical text)

(Krishna, Wieting, and Iyyer 2020). XFORMAL, a formality-based dataset, includes text in several languages, including Portuguese, French, and Italian (Briakou et al. 2021).

Reference-free frameworks that focus on various textual features (such as readability-based ones) have also been applied to style transfer evaluation, an example being Coh-Metrix (McNamara et al. 2014). Perplexity based on large language models has also been utilised as a measure of fluency (Jin et al. 2022).

6.4 Takeaways

Unlike the previously discussed tasks of simplification and summarisation, style transfer implies change of a more subjective and qualitative nature, thus providing a strong link with literary adaptation. In addition, the newly-found focus on the reader as well as many of the textual features evoked as representative of style are directly related to the adaptation of literary texts in view of different audience needs and preferences: reader profiles, authorship, modernisation, and even sentiment polarity (e.g. a more "positive" text may be directed at a younger audience). Extending García et al. (2021)'s reasoning, we could unify a large number of both text-centred and reader-centred textual variables, including even language, under a common term tentatively referred to as "style".

As demonstrated throughout the current section, isolated evaluation metrics and limited in size and breadth datasets cannot easily capture the nature of style transfer as a multifaceted framework. Prompt engineering of state-of-the-art LLMs constitutes a promising direction; however, the technique is associated with low explainability and replicability as well as a lack of clarity in terms of the definition of style. A larger yet carefully categorised selection of independent evaluation metrics would likely be of benefit to the evaluation of literary adaptation as a crossroads of textual styles. Ideally, human evaluation including comprehensive studies of reader experience, would also be included.

7. Related NLP Tasks: Creative Generation

7.1 Definition

Creativity is typically associated with human intelligence and with notions such as novelty, personality, uniqueness, emotion and aesthetics. In her work on Artificial Intelligence and creativity, Boden (1998) differentiates between psychological and historical creativity, where the former is limited to a single person and the latter extends to humanity as a whole. She also defines combinational, exploratory, and transformational creativity, which move incrementally from reuse to actual novelty. Although the notion of creative generation (and any creativity for that matter) might strike as an antithesis of technology, NLP methods have been applied at different steps of the process, and current generative models are capable of producing full texts of different genres and sizes, such as short stories and poems, that largely mimic human creative writing.

7.2 Timeline

In 1950, the Turing Test, whose purpose was to evaluate machines by proxy of their resemblance to humans in linguistic interaction, sparked interest in the potential of technological tools to perform creative tasks, such as storytelling (Turing 1950). The "rule-based" NLP period saw the introduction of simple creative tasks, such as the production of pun-like jokes by JAPE (Joke Analysis and Production Engine) (Binsted 1996). The underlying mechanism included schemas that define the structure of jokes, templates of realisation, lexical resources and estimation of phonological similarity.

Real breakthrough came with Transformer models and sequence-to-sequence text generation in the late 2010s. A state-of-the-art convolutional model of story writing was proposed by Fan, Lewis, and Dauphin (2018). It made use of a hierarchical framework: firstly, a single sentence (referred to as "prompt") that describes a story was generated, based on which further generation was conditioned, guaranteeing consistency and staying on-topic. A dataset of human-made stories and related prompts was used in training the model. A year later, a GPT-2 model outperformed Fan, Lewis, and Dauphin (2018)'s model in the following aspects: order of events, lexical variety and reliance on the provided prompt. Both models were noted to exhibit the shortcoming of excessive repetitiveness. Also in 2019, Trăuşan-Matu claimed that despite recent advances, stories generated by language models still lacked in empathy and human-like creativity. In his survey of NLG and creative generation that covers the years 2016 to 2021, Alsharhan (2022) identified 16 high-quality papers, 75% of which evaluated positively the impact of current technology on creative writing. Shanahan and Clarke (2023) used prompting strategies and temperature fine-tuning of GPT-4 in a task of story writing and evaluated the output qualitatively, reaching the conclusion that LLMs can competitively engage in creative writing, while one should keep in mind their currently high dependence on user prompts. Gómez-Rodríguez and Williams (2023) compared the creative generation abilities of several LLMs to each other as well as to a human benchmark for the same task, applying human evaluation to the derived stories, also coming to mixed conclusions. Despite performing very highly, commercial LLMs were judged not to match human writers in terms of originality. Specifically, humour was concluded to be an emerging ability of LLMs.

The implication of LLMs in creative writing gave rise to questions of ethical and philosophical nature. Referring to Boden's notions of value, novelty and surprise, Franceschelli and Musolesi (2024) claimed that the creativity of LLMs is, by definition and insurmountably, limited in both nature and scope. What about the fair use of LLMs? Should one apply limitations to their ability to impersonate an author's style? Can an LLM own authorship? Do the authors of texts that the model was trained on share in its authorship? Currently, there are publishers and literary contest organisers that ask authors to fill in declarations about potential use of AI in their work, although no global frameworks on the topic have been established (Coeckelbergh 2020).

The use of contemporary AI tools in discrete modular tasks related to creative writing (such as character development and plot outlining) is typically seen as an ethically safe practice and as an example of effective human-machine cooperation. Kreminski and Martens (2022) offer an overview of LLMs' potential to support creative writers in cases of "writer's

block". The tool Story Centaur makes use of LLMs' few-shot abilities at discrete writing tasks, such as the generation of a sentence based on the previous one and a "magic word" provided by the user (Swanson et al. 2021).

Additional NLP tasks that imply an element of creative generation have been considered challenging in terms of implementation and evaluation. Toral and Way (2018, p. 263) referred to the translation of literary texts as "the greatest challenge for [pre-LLM] M[achine] T[ranslation]". At a 2023 shared task on discourse-level literary translation organised by Tencent AI Lab and China Literature Ltd, LLMs were the clear winner based on both human and automatic evaluation (Wang et al. 2023b).

7.3 Evaluation

Fan, Lewis, and Dauphin (2018)'s sequence-to-sequence model made use of measures of perplexity on the test set as well as of prompt ranking accuracy in its evaluation of specific aspects of creative writing, such as "staying on topic". Given the task of creative writing translation, Wang et al. (2023b) opted for document-level SacreBLEU[12] (d-BLEU). The limitations of quantitative evaluation of automatic creative output become more prominent with the technology's advancement and a movement away from lower-level concerns such as those of grammaticality, fluency, and cohesion. Boden (1998) notes that evaluation of high-level creativity is largely impossible, as the implied method would, by definition, need to be external to previously established frameworks in order to be applicable to true novelty. The following are examples of human-evaluation-based measures that have been associated with automatic creative generation: characterisation, imagery, humour, and use of idioms. Their quantification and systematisation, however, remains challenging.

7.4 Takeaways

The current quality of automatic tools, notably LLMs, in relation to creative writing, give reasons to be optimistic about NLP's readiness for literary adaptation. However, whilst language proficiency and fluency are unlikely to come as a significant concern, the situation might be different in relation to the knowledge and application of diverse artistic traditions as associated with different cultures.

Key limitations to automatic creative generation, such as ethical concerns related to authorship and the lack of objectivity at evaluation, apply to a lesser extent to the practice of literary adaptation due to the presence of an original text that the output may be compared against. At the same time, additional questions may emerge that require careful contemplation and possibly regulations. For instance: should there be a limit of the extent to which the original text may be reused? Does the original author's copyright come into play?

Due to the presence of an original text, comparison-based evaluation measures such as BLEU score may be relevant to literary adaptation, while likely insufficient on their

12 a standardised implementation of BLEU

own. Elements of creative writing that have been applied in human evaluation of machine-produced text, such as humour, imagery, and idioms, are also to be incorporated. It is worth noting that as semantics is steadily making its way into NLP tools and methods, textual features that were once seen as strictly qualitative, such as the presence of metaphors in text, are currently identifiable and measurable by neural models (Wachowiak, Gromann, and Xu 2022).

8. Discussion

8.1 General Trends

The perceived lagging behind of literary adaptation within NLP in comparison to other tasks may be understandable given its less practical nature and lower sense of urgency compared to, for instance, work such as Segura-Bedmar and Martínez (2017)'s simplification of drug package texts. However, in terms of technological readiness, automated literary adaptation can be accommodated as of today. Common NLP trends of relevance include the strong linguistic, including multilingual, performance of LLMs, a growing focus on style and qualitative textual characteristics, and reader-centredness. One may also notice that as technology advances, user groups tend to be more narrowly defined, an example being readers with different language proficiency levels. Also, associations of two or more discrete tasks (such as translation and style transfer) become more common, and literary adaptation in itself is, as discussed, a meeting point of several practices that are prone to automation, such as simplification, summarisation, style change and literary production. Finally, although the current success of creative text generation is ambiguous, literary adaptation largely mitigates its limitations by being based on a source text.

In terms of the exact generation methods that literary adaptation is most likely to benefit from, LLMs are the first candidate given the current context in NLG, and their performance may be efficiently complemented with established techniques such as prompt engineering and RAG. Additional quantitative and qualitative development within NLP in the near future is likely to continue paving the way for literary adaptation; in particular, in relation to its concerns with textual length, coherence, plot, and figurative language.

8.2 Evaluation Framework

One of the main challenges of automatic literary adaptation is the need to elaborate a relevant evaluation framework. See Appendix A for a table presenting the automatic metrics mentioned within the current survey, along with their broad types, applications and characteristics/benefits. These metrics may be roughly categorised into surface-based, semantic, learned, and reference-free. Surface-level metrics such as BLEU and readability scores are still widely used due to their simplicity combined with the ability to capture key textual characteristics. Although they typically come with significant limitations (such as BLEU's inability to account for paraphrased text), these can be alleviated when multiple metrics are used jointly in carefully selected combinations.

The combination of a number of different automatic metrics can help construct a reusable and adjustable evaluation framework for automatically generated text, in particular in relation to the task of literary adaptation. Additional metrics that we propose to be included in the framework by merit of their relevance to literary adaptation as well as additional tasks addressed in this survey include: Biber (1988)'s comprehensive selection of textual features (including, but not limited to, morphology- and syntax-related ones), literature-specific features, such as figurative language, humour and surprise, as currently measurable by neural tools (e.g. Wachowiak, Gromann, and Xu 2022; Fan et al. 2020) and different aggregators for applicable features, such as average, minimal and maximal values, which have also proved relevant in the description of a text's linguistic qualities (Wilkens et al. 2022).

Many of the mentioned automatic metrics require reference texts. However, reliance on multiple limited datasets in relation to different audience types and languages would be highly inefficient. Instead, we propose the composition of a large literary corpus and the establishment of gold standard values for the various utilised metrics in relation to it. Ideally, the corpus should contain a large number of paired original and professionally adapted versions of the same text, possibly along with additional original texts that account for lower-resource settings, thereby helping include a variety of genres, languages and literary traditions. Based on estimation of feature relevance with respect to different audience types and languages, as well as on correlation analyses that can help efficiently reduce the number of utilised features, different evaluation formulas may be defined for different adaptation scenarios. Finally, the gathered insight about feature relevance may be used not only in the context of evaluation but also in automatic editing of output text or as a protocol integrated into LLM prompts.

8.3 Human Involvement

Currently, despite being costly in terms of time, effort and money, human-based evaluation is the preferred standard in all discussed NLP tasks. The need for substantial human participation is reinforced by concerns of an ethical and philosophical nature in relation to the role of professionals, such as when it comes to creative output and to pedagogical applications.

It is best practice for automated metrics to be validated through their correlation with human evaluation, as demonstrated by Vajjala and Lucic (2019). In the specific case of literary adaptation, a selected reader audience may be asked both to answer comprehension questions, a practice shown to directly measure meaning preservation (Agrawal and Carpuat 2024), and to subjectively evaluate their reading experience and appreciation of the text.

Also, the task of literary adaptation task can realistically be seen as partially rather than fully automatable. Involvement in the process of professionals in fields such as foreign language teaching would be crucial and extend from prompt composition and use to manual editing of automatically produced text.

9. Conclusion

Whilst the independent generation of high-quality creative text by NLP tools is not a current reality, concepts relevant to literature, such as humour, constitute emergent abilities of contemporary models. Given the advanced landscapes in relation to the relevant tasks of automatic text simplification and summarisation as well as style transfer and creative generation, it is not far-fetched to envision the beginning of systematic automated literary adaptation. This survey lays out methods and challenges that the task is likely to be met with and that are to be kept in mind in view of a process of preparation as well as the initial implementation steps of the emerging task.

The immediate steps that we recommend be taken in view of the proposed task include the compilation of a dedicated literary adaptation corpus of a substantial size and the definition of reference-free formulas (similar to readability ones) to be used in the evaluation of adapted literary texts. The corpus should consist of pairs of original and adapted literary texts as annotated according to fine-grained audience categories (e.g. age ranges for children and proficiency levels for foreign language learners). Importantly, a variety of languages should be represented within the corpus. A statistical analysis should then be applied to the corpus, based on a variety of quantitative metrics as laid out in Section 8.2. The most relevant (yet not strongly correlated) metrics should be determined by audience and by language, and corresponding formulas should be formulated for future use. The described framework would pave the way for standardised adaptation effort that extends to a variety of audiences and accounts for a diversity of literary traditions.

10. Limitations

Relevance to literary adaptation is not reserved to the four NLP tasks that we chose to focus on. For instance, machine translation would be heavily implied in cases of different source and target languages. Another task, paraphrase, is also relevant in its transformation of textual formulation whilst meaning is preserved. However, due to the paraphrase's under-specified nature, we have opted for the inclusion of tasks that are associated with concrete communicative objectives that are typically shared with those of literary adaptation: reduction in size (in summarisation), reduction in complexity (in simplification) and change in style to appeal aesthetically to an alternative e.g. more modern audience (style transfer).

Also, we should note that the current survey draws conclusions about the potential of a yet undeveloped NLP task, namely that of literary adaptation, based on related tasks. The nature of the intersection of the described tasks may be hypothesised but is bound to lack certainty. It is therefore important that the proposed guidelines for literary adaptation be factually experimented with. Finally, the mentioned guidelines are of a technical and scientific rather than ethical nature. The ethical aspects and associated potential limitations of automatic literary adaptation, which include for instance questions of access and authorship, have only slightly been touched upon and go beyond the scope of this study.

References

Abdel-Salam, Shehab and Ahmed Rafea. 2022. Performance study on extractive text summarization using BERT models. *Information*, 13(2).

Agrawal, Sweta and Marine Carpuat. 2024. Do text simplification systems preserve meaning? A human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Al-Thanyyan, Suha S. and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys*, 54(2):Article 43, 1–36.

Alsharhan, Abdulla. 2022. Natural language generation and creative writing: A systematic review. *International Journal of Advances in Applied Computational Intelligence*, 1(1):69–90.

Aluísio, Sandra Maria and Caroline Gasperin. 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 (YIWCALA '10), Los Angeles, CA, USA*, page 46–53, Association for Computational Linguistics.

Alva-Manchego, Fernando, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Association for Computational Linguistics.

Alva-Manchego, Fernando, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China: System Demonstrations*, pages 49–54, Association for Computational Linguistics.

Aranzabe, María, Arantza Ilarraza, and Itziar Gonzalez-Dios. 2012. First approach to automatic text simplification in Basque. In *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop*, pages 1–8.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan*, pages 65–72, Association for Computational Linguistics.

Benedetto, Luca and Paula Buttery. 2025. Towards CEFR-targeted text simplification for question adaptation. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing – Natural Language Processing in the Generative AI Era, Varna, Bulgaria*, pages 150–157, INCOMA Ltd., Shoumen, Bulgaria.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge. Reprinted/online publication: 2012.

Binsted, Kim. 1996. *Machine Humour: An Implemented Model of Puns*. Ph.D. thesis, University of Edinburgh.

Biran, Or, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: A context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Association for Computational Linguistics.

Boden, Margaret A. 1998. Creativity and artificial intelligence. *Artificial Intelligence*, 103(1-2):347–356.

Bonifacci, Paola, Cinzia Viroli, Chiara Vassura, Elisa Colombini, and Lorenzo Desideri. 2023. The relationship between mind wandering and reading comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, 30(1):40–59.

Briakou, Eleftheria, Di Lu, Ximing Lu, and Joel Tetreault. 2021. Olá, Bonjour, Salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Association for Computational Linguistics.

Brunato, Dominique, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of the 9th Linguistic Annotation Workshop, Denver, Colorado, USA*, pages 31–41, Association for Computational Linguistics.

Caplan, David. 2012. *Language: Structure, Processing, and Disorders.* MIT Press, Cambridge, MA.
Cersosimo, Rita, Filippo Domaneschi, and Alice Cancer. 2025. The impact of metaphors on academic text comprehension: The case of students with dyslexia. *Dyslexia*, 31(1).
Chang, Yapei, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. ArXiv, abs/2310.00785.
Chen, Han-Bin, Hen-Hsen Huang, Hsin-Hsi Chen, and Ching-Ting Tan. 2012. A simplification-translation-restoration framework for cross-domain SMT applications. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India*, pages 545–560.
Chen, Kai, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Improving transformer-based neural machine translation with recurrent and convolutional components. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 195–206.
Cheng, Jianpeng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany (Volume 1: Long Papers)*, pages 484–494, Association for Computational Linguistics.
Chopra, Sumit, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, Association for Computational Linguistics.
Coeckelbergh, Mark. 2020. *AI Ethics.* MIT Press, Cambridge, MA.
Coster, William and David Kauchak. 2011a. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation, Portland, Oregon, USA*, pages 1–9, Association for Computational Linguistics.
Coster, William and David Kauchak. 2011b. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Association for Computational Linguistics.
Cripwell, Liam, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia*, pages 993–1006, Association for Computational Linguistics.
Day, Richard R. and Julian Bamford. 1998. *Extensive Reading in the Second Language Classroom.* Cambridge University Press, Cambridge.
Devlin, Siobhan. 1999. *Simplifying Natural Language Text for Aphasic Readers.* Ph.D. thesis, University of Sunderland, Sunderland, United Kingdom.
Devlin, Siobhan and John Tait. The use of a psycholinguistic database in the simplification of text for aphasic readers. In J. Nerbonne, editor, *Linguistic Databases.* CSLI Publications, Stanford, California, pages 161–173.
Devlin, Siobhan and Gary Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '06), Portland, Oregon, USA*, pages 225–226, ACM.
Dickey, Michael Walsh and Cynthia K. Thompson. 2009. Automatic processing of wh- and NP-movement in agrammatic aphasia: Evidence from eyetracking. *Journal of Neurolinguistics*, 22(5):563–584.
Dickinson, Paul. 2017. Effects of extensive reading on EFL learner reading attitudes. In *Proceedings of the 21st Conference of the Pan-Pacific Association of Applied Linguistics, Tamkang University, Taiwan*, pages 28–35.
Dmitrieva, Anna and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, Kiyv, Ukraine*, pages 73–79, Association for Computational Linguistics.
Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT 2002)*, pages 138–145, Morgan Kaufmann.
Douma, Willem Hendrik. 1960. *De Leesbaarheid van Landbouwbladen: Een Onderzoek naar en een Toepassing van Leesbaarheidsformules.* Veenman & Zonen, Wageningen, Netherlands.
DuBay, William H. 2004. *The Principles of Readability.* Impact Information, Costa Mesa, CA, USA.

El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Ellis, John. 1982. The literary adaptation. *Screen*, 23(3):3–5.

Evans, Richard, Constantin Orăsan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014) Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 131–140, Association for Computational Linguistics.

Ewers, Hans-Heino. 2009. *Fundamental Concepts of Children's Literature Research: Literary and Sociological Approaches.* Routledge, London.

Fan, Angela, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia (Volume 1: Long Papers)*, pages 889–898, Association for Computational Linguistics.

Fan, Xiaochao, Hongfei Lin, Liang Yang, Yufeng Diao, Chen Shen, Yonghe Chu, and Yanbo Zou. 2020. Humor detection via an internal and external neural network. *Neurocomputing*, 394:105–111.

Farajidizaji, Asma, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? An initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia*, pages 9325–9339, ELRA and ICCL.

Feng, Yutao, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. ArXiv, abs/2302.11957.

FitzGerald, Elizabeth, Ann Jones, Natalia Kucirkova, and Eileen Scanlon. 2018. A literature synthesis of personalised technology-enhanced learning: What works and why. *Research in Learning Technology*, 26:2095.

Franceschelli, Giorgio and Mirco Musolesi. 2024. On the creativity of large language models. *AI & Society*, 40(5):3785–3795.

Freyer, Nils, Hendrik Kempt, and Lars Klöser. 2024. Easy-Read and large language models: On the ethical dimensions of LLM-based text simplification. *Ethics and Information Technology*, 26(3):1–10.

Fu, Zhenxin, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: exploration and evaluation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*, AAAI Press.

Gala, Núria, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France*, pages 1353–1361, European Language Resources Association.

García, Xavier, Noah Constant, Mandy Guo, and Orhan Firat. 2021. Towards universality in multilingual text rewriting. ArXiv, abs/2107.14749.

Genest, Pierre-Etienne and Guy Lapalme. 2012. Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea (Volume 2: Short Papers)*, pages 354–358, Association for Computational Linguistics.

Glavaš, Goran and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China (Volume 2: Short Papers)*, pages 63–68, Association for Computational Linguistics.

Godwin-Jones, Robert. 2018. Contextualized vocabulary learning. *Language Learning & Technology*, 22(3):1–19.

Gómez-Rodríguez, Carlos and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore*, pages 14504–14528, Association for Computational Linguistics.

Han, Jingxuan, Quan Wang, Zikang Guo, Benfeng Xu, Licheng Zhang, and Zhendong Mao. 2024. Disentangled learning with synthetic parallel data for text style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand (Volume 1: Long Papers)*, pages 15187–15201, Association for Computational Linguistics.

Happé, Francesca G. E. 1993. Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2):101–119.

Haverals, Wouter, Lindsey Geybels, and Vanessa Joosen. 2022. A style for every age: A stylometric inquiry into crosswriters for children, adolescents and adults. *Language and Literature*, 31(1):62–84.

Hill, David R. 2013. Graded readers. *ELT Journal*, 67(1):85–125.

Hovy, Eduard, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*, European Language Resources Association (ELRA).

Hu, Baotian, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1967–1972.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, Association for Computing Machinery, New York, NY, USA.

Huang, Lei, Yanxiang He, Furu Wei, and Wenjie Li. 2010. Modeling document summarization as multi-objective optimization. In *Proceedings of the Third International Symposium on Intelligent Information Technology and Security Informatics*, pages 382–386.

Hutcheon, Linda. 2006. *A Theory of Adaptation*. Routledge, New York.

Huynh, Linh and Danielle S. McNamara. 2025. GenAI-powered text personalization: Natural language processing validation of adaptation capabilities. *Applied Sciences*, 15(12).

Intrigi.bg. 2024. Adaptatsiyata na Rusi Chanev na "Pod igoto" predizvika skandal. Published: December 7, 2024.

Jain, Sameer, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada*, pages 8487–8495, Association for Computational Linguistics.

Jamet, Henri, Yash Raj Shrestha, and Michalis Vlachos. 2024. Difficulty estimation and simplification of French text using LLMs. In *Generative Intelligence and Intelligent Tutoring Systems. Proceedings of the 20th International Conference ITS 2024, Thessaloniki, Greece*, pages 395–404, Springer Nature Switzerland.

Jhamtani, Harsh, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark*, pages 10–19, Association for Computational Linguistics.

Jin, Di, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Kalandadze, Tamar, Courtenay Frazier Norbury, Terje Nærland, and Kari-Anne B. Næss. 2018. Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism*, 22(2):99–117.

Kazantseva, Anna and Stan Szpakowicz. 2010. Summarizing short stories. *Computational Linguistics*, 36(1):71–109.

Kew, Tannon, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore*, pages 13291–13309, Association for Computational Linguistics.

Klaper, David, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2013), Sofia, Bulgaria*, pages 11–19, Association for Computational Linguistics.

Ko, Youngjoong and Jungyun Seo. 2008. An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. *Pattern Recognition Letters*, 29(9):1366–1371.

Kobayashi, Hayato, Masaki Noguchi, and Taichi Yatsuka. 2015. Summarization based on embedding distributions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*

Processing, Lisbon, Portugal, pages 1984–1989, Association for Computational Linguistics.

Krashen, Stephen D. 1982. *Principles and Practice in Second Language Acquisition*. Pergamon Press, Oxford.

Kreminski, Max and Chris Martens. 2022. Unmet creativity support needs in computationally supported creative writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022), Dublin, Ireland*, pages 74–82, Association for Computational Linguistics.

Krishna, Kalpesh, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Ladhak, Faisal, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048.

Lefebvre, Julie. 2021. "Adaptation" ou "texte intégral" ? Représentations éditoriales de l'enfant lecteur. *Le Français Aujourd'hui*, 213(2):53–64.

Li, Juncen, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, Association for Computational Linguistics.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Association for Computational Linguistics.

Liu, Pusheng, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia*, pages 15285–15295, ELRA and ICCL.

Liu, Yan. 2021. Readability and adaptation of children's literary works from the perspective of ideational grammatical metaphor. *Journal of World Languages*, 7(2):334–354.

Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Maddela, Mounica, Yao Dou, David Heineman, and Wei Xu. 2022. LENS: A learnable evaluation metric for text simplification. ArXiv, abs/2212.09739.

Mar, Raymond A., Jingyuan Li, Anh T. P. Nguyen, and Cindy P. Ta. 2021. Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review*, 28(3):732–749.

Mar, Raymond A., Keith Oatley, Jacob B. Hirsh, Jennifer dela Paz, and Jordan B. Peterson. 2006. Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality*, 40(5):694–712.

Martinussen, Rhonda, Josephine Hayden, Sheilah Hogg-Johnson, and Rosemary Tannock. 2005. A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44(4):377–384.

Mashal, Nira and Anat Kasirer. 2012. Principal component analysis study of visual and verbal metaphoric comprehension in children with autism and learning disabilities. *Research in Developmental Disabilities*, 33(1):274–282.

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Melogno, Sergio, Maria A. Pinto, and Margherita Orsolini. 2017. Novel metaphors comprehension in a child with high-functioning autism spectrum disorder: A study on assessment and treatment. *Frontiers in Psychology*, 7:2004.

Mizukami, Masahiro, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Linguistic individuality transformation for spoken language. In G.G. Lee, H.K. Kim, M. Jeong, and J.-H. Kim, editors, *Natural Language Dialog Systems and Intelligent Assistants*. Springer International Publishing, Cham, pages 129–143.

Mukherjee, Sourabrata, Vojtěch Hudeček, and Ondřej Dušek. 2023. Polite chatbot: A text style transfer application. In *Proceedings of the 17th Conference of the European Chapter of the Association for*

Computational Linguistics, Dubrovnik, Croatia: Student Research Workshop, pages 87–93, Association for Computational Linguistics.

Müller, Anja, editor. 2013. *Adapting Canonical Texts in Children's Literature.* Bloomsbury Academic, London.

Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)@NAACL-HLT 2012, Montrèal, Canada*, pages 95–100.

Nation, I. S. Paul. 2001. *Learning Vocabulary in Another Language.* Cambridge University Press, Cambridge.

National Institute of Standards and Technology. 2004. Document Understanding Conference (DUC). https://duc.nist.gov.

Nenkova, Ani, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4–es.

Nières-Chevrel, Isabelle. 2009. *Introduction à la littérature de jeunesse.* Didier Jeunesse, Paris.

Nisioi, Sergiu, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada (Volume 2: Short Papers)*, pages 85–91, Association for Computational Linguistics.

Nussbaum, Martha Craven. 2010. *Not for Profit: Why Democracy Needs the Humanities.* Princeton University Press, Princeton, NJ.

Oatley, Keith. 2011. *Such Stuff as Dreams: The Psychology of Fiction.* Wiley-Blackwell, Chichester, UK.

van Oosten, Philip, Dieter Tanghe, and Véronique Hoste. 2010. Towards an improved methodology for automated readability prediction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta*, European Language Resources Association (ELRA).

Ormaechea, Lucía and Nikos Tsourakis. 2024. Automatic text simplification for French: Model fine-tuning for simplicity assessment and simpler text generation. *International Journal of Speech Technology*, 27:957–976.

Ouyang, You, Wenjie Li, Sujian Li, and Qin Lu. 2011. Applying regression models to query-focused multi-document summarization. *Information Processing & Management*, 47(2):227–237.

Owczarzak, Karolina. 2009. DEPEVAL(summ): Dependency-based evaluation for automatic summaries. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore*, pages 190–198.

Paetzold, Gustavo and Lucia Specia. 2016. SV000gg at SemEval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California*, pages 969–974, Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA*, pages 311–318, Association for Computational Linguistics.

Pu, Xiao, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. ArXiv, abs/2309.09558.

Radev, Dragomir R., Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

Rao, Sudha and Joel Tetreault. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 129–140, Association for Computational Linguistics.

Reddy, Sravana and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science, Austin, Texas*, pages 17–26, Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Association for Computational Linguistics.

Reif, Emily, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland (Volume 2: Short Papers)*, pages 837–848, Association for Computational Linguistics.

Restrepo Ramos, Fabian Dario. 2015. Incidental vocabulary learning in second language acquisition: A literature review. *PROFILE Issues in Teachers' Professional Development*, 17(1):157–166.

Rios, Annette, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization, Dominican Republic*, pages 152–161, Association for Computational Linguistics.

Wan-a rom, Uthai. 2008. Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language*, 20(1):43–69.

Rutherford, Marion, Julie Baxter, Zoe Grayson, Lorna Johnston, and Anne O'Hare. 2020. Visual supports at home and in the community for individuals with autism spectrum disorder. *Autism*, 24(2):447–457.

Ryan, Michael J., Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada (Volume 1: Long Papers)*, pages 4898–4927, Association for Computational Linguistics.

Scialom, Thomas, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021a. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic*, pages 6594–6604, Association for Computational Linguistics.

Scialom, Thomas, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021b. Rethinking automatic evaluation in sentence simplification. ArXiv, abs/2104.07560.

Segura-Bedmar, Isabel and Paloma Martínez. 2017. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of Biomedical Semantics*, 8(1):45.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Association for Computational Linguistics.

Shanahan, Murray and Catherine A. M. Clarke. 2023. Evaluating large language model creativity from a literary perspective. ArXiv, abs/2312.03746.

Shardlow, Matthew, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 Shared Task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Mexico City, Mexico*, pages 571–589, Association for Computational Linguistics.

Shavit, Zohar. 1999. The double attribution of texts for children and how it affects writing for children. In Sandra L. Beckett, editor, *Transcending Boundaries: Writing for a Dual Audience of Children and Adults*. Garland Publishing, New York and London, pages 83–98.

Shen, Tianxiao, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*.

Shetty, Krithi and Jagadish S. Kallimani. 2017. Automatic extractive text summarization using k-means clustering. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 1–9.

Snover, Matthew, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? exploring different human judgments with a tunable MT metric. In *Proceedings of the*

Fourth Workshop on Statistical Machine Translation (WMT 2009), Athens, Greece, pages 259–268, Association for Computational Linguistics.

Snowling, Margaret J., Charles Hulme, and Kate Nation. 2000. Defining and understanding dyslexia: Past, present and future. *Oxford Review of Education*, 46(4):501–513.

Spaulding, Seth. 1956. A spanish readability formula. *The Modern Language Journal*, 40(7):433–441.

Stajner, Sanja and Horacio Saggion. 2013. Adapting text simplification decisions to different text genres and target users. *Procesamiento del Lenguaje Natural*, 51.

Štajner, Sanja and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan*, pages 374–382, Asian Federation of Natural Language Processing.

Stodden, Regina. 2024. EASSE-DE & EASSE-multi: Easier automatic sentence simplification evaluation for german & multiple languages. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Miami, Florida, USA*, pages 107–116, Association for Computational Linguistics.

Stymne, Sara, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), Oslo, Norway*, pages 375–386, Linköping University Electronic Press.

Sulem, Elior, Omri Abend, and Ari Rappoport. 2018. 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, New Orleans, Louisiana, Volume 1 (Long Papers)*, pages 685–696, Association for Computational Linguistics.

Surya, Sai, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*, pages 2058–2068, Association for Computational Linguistics.

Swanson, Ben, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story Centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256, Association for Computational Linguistics.

Thiel, Elizabeth. 2013. Downsizing Dickens: Adaptations of Oliver Twist for the child reader. In Anja Müller, editor, *Adapting Canonical Texts in Children's Literature*. Bloomsbury Academic, London, pages 143–162.

Tikhonov, Alexey and Ivan P. Yamshchikov. 2018. What is wrong with style transfer for texts? ArXiv, abs/1808.04365.

Toral, Antonio and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment: From Principles to Practice*. Springer, pages 263–287.

Tratz, Stephen and Eduard Hovy. 2008. Summarization evaluation using transformed basic elements. In *Proceedings of the First Text Analysis Conference (TAC 2008), National Institute of Standards and Technology Gaithersburg, Maryland, USA*.

Trăuşan-Matu, Ştefan. 2019. Computer-based story generation: An analysis from a phenomenological standpoint. *International Journal of User-System Interaction*, 12(1):39–53.

Turing, Alan. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Vajjala, Sowmya and Ivana Lučić. 2018. OneStopEnglish Corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans, Louisiana*, pages 297–304, Association for Computational Linguistics.

Vajjala, Sowmya and Ivana Lucic. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of Natural Language Processing for Building Educational Applications, Florence, Italy*, pages 349–359, Association for Computational Linguistics.

Van Lierop-Debrauwer, Helma. 2000. Over de 'Grote Gelijkenis': adolescentenromans voor jongeren en voor volwassenen. *Literatuur Zonder Leeftijd*, 14:336–351.

Vazov, Ivan Minchov. 2024. *Pod igoto*. Chanev, Rusi (ed.). Iztok-Zapad, Sofia. Adapted edition.

Wachowiak, Lennart, Dagmar Gromann, and Chao Xu. 2022. Drum up support: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53.

Wang, Jiaan, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th Workshop on New Frontiers in Summarization (NewSum 2023), Toronto, Canada*, pages 12–23, Association for Computational Linguistics.

Wang, Longyue, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 Shared Task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)*.

Wang, Yingjie and Jun Ma. 2013. A comprehensive method for text summarization based on latent semantic analysis. In *Proceedings of the Natural Language Processing and Chinese Computing: Second CCF Conference (NLPCC 2013), Chongqing, China*, pages 394–401, Springer, Berlin, Heidelberg.

Washburne, Carleton and Mabel Vogel. 1926. *Winnetka Graded Book List*. American Library Association, Chicago, Illinois.

Wilkens, Rodrigo, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey, and Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France*, pages 1217–1233, European Language Resources Association.

Wu, Yunshu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, St. Julian's, Malta (Volume 2: Short Papers)*, pages 330–343, Association for Computational Linguistics.

Xiao, Le and Xiaolin Chen. 2023. Enhancing LLM with evolutionary fine tuning for news summary generation. ArXiv, abs/2307.02839.

Xu, Wei. 2017. From Shakespeare to Twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark*, pages 1–9, Association for Computational Linguistics.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xu, Wei, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xu, Wei, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India*, pages 2899–2914.

Yang, Ziyu. 2024. *Enhancing the Comprehension: Text Simplification Approaches and the Role of Large Language Models*. Phd dissertation, Temple University.

Yuan, Weizhe, Graham Neubig, and Pengfei Liu. 2021. BARTSCORE: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Curran Associates Inc., Red Hook, NY, USA.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, OpenReview.net.

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, MIT Press, Cambridge, MA, USA.

Zhang, Xingxing and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark*, pages 584–594, Association for Computational Linguistics.

Zhang, Yang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2026. A comprehensive survey on automatic text summarization with exploration of LLM-based methods. *Neurocomputing*, 663:131928.

Zhao, Junbo (Jake), Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. ArXiv, abs/1706.04223.

Appendix

A. Evaluation Metrics

Metric	Type	Application	Benefits/Characteristics
BLEU	n-gram overlap	summarisation, simplification, style transfer, creative generation	widely used; good for precision; simple to compute; multiple variants
SacreBLEU	n-gram overlap	summarisation, simplification, creative generation	reproducible; standardised;
ROUGE	n-gram overlap	summarisation, simplification, style transfer	good for recall; widely used; multiple variants
NIST	weighted n-gram overlap	simplification	accounts for informativeness; improves over BLEU
TERp	edit distance	simplification	interpretable; reference-based
METEOR	n-gram + semantic matching	summarisation, simplification, style transfer	accounts for synonyms; considers word order; better correlation than BLEU
SARI	operation-based (add/delete/keep)	simplification	task-specific; captures simplicity; interpretable
SAMSA	semantic structure	simplification	reference-free; captures semantics; focuses on predicate-argument relations
BERTScore	semantic similarity (embedding-based)	summarisation, simplification, style transfer	captures semantics; robust to paraphrasing; widely used

Metric	Type	Application	Benefits/Characteristics
COMET	learned metric	simplification	models human judgement; multilingual; uses source and reference
BLEURT	learned metric (BERT-based)	summarisation	models human judgement; robust with limited references
BARTScore	LLM-based (seq2seq scoring)	summarisation	correlates with fluency; generation-based evaluation
QuestEval	QA-based	summarisation, simplification	evaluates factual consistency; reference-free; correlates with human judgement
LENS	learned metric	simplification	correlates with fluency, meaning, simplicity; high correlation with humans
FKBLEU	hybrid (BLEU + readability)	simplification	combines similarity and simplicity; uses source and reference
Flesch Reading Ease	readability	simplification, style transfer	simple to compute; reference-free
Flesch–Kincaid Grade Level (FKGL)	readability	simplification, style transfer	interpretable; grade-level estimate; widely used
Perplexity	language model-based	style transfer, creative generation	correlates with fluency; reference-free

Table 1
Automatic evaluation metrics mentioned within the survey as relevant to the NLP tasks of simplification, summarisation, style transfer and creative generation. Individual shortcomings are not focused on, as the purpose is to provide an overview of different methods' strengths that allow for their effective use when applied in combinations rather than independently.