Novella Tedesco*, Silvia Bernardini and Federico Garcea

English as a lingua franca in academic publishing: using round-trip translation to estimate linguistic revision difficulty

https://doi.org/10.1515/jelf-2025-2002

Abstract: In this paper the feasibility of a novel method for estimating revision difficulty of academic texts to be published in English as a lingua franca (ELF) is examined with the purpose of optimizing revisors' workflow and performance. In doing so, we also address the complex issue of linguistic standards in ELF academic productions from a new perspective. The method of automatic round-trip translation (RTT), which consists in machine-translating a text into another language and back into the source language, is applied in conjunction with BLEU, an automatic text similarity method traditionally used in machine translation evaluation. The similarity between a manuscript and its round-trip translation is calculated to obtain a priori evaluation of revision difficulty, and the efficacy and reliability of the method are tested. Alongside with the experimental research, some theoretical issues regarding ELF and the practice of revision in the hyperglobalized context of international academic publishing are discussed. The results of the experiments show that the RTT-based method can be successfully used to estimate revision difficulty. The variety of data used in the experiments demonstrates the applicability of the method beyond the cases under study. Some limitations and further developments of the study are also briefly discussed.

Keywords: BLEU; English as a lingua franca; English for academic purposes; revision; round-trip translation

Astratta: In questo articolo viene presentato un metodo innovativo per stimare la difficoltà di revisione di testi accademici da pubblicare in inglese come lingua franca, allo scopo di ottimizzare il lavoro e le performance dei revisori. In questo modo la complessa questione degli standard linguistici nella produzione accademica in

***Corresponding author: Novella Tedesco**, Department of Translation and Interpreting, University of Bologna, Bologna, Italy, E-mail: novella.tedesco2@unibo.it
Silvia Bernardini and Federico Garcea, Department of Translation and Interpreting, University of Bologna, Bologna, Italy, E-mail: silvia.bernardini@unibo.it (S. Bernardini), federico.garcea2@unibo.it (F. Garcea)

inglese viene affrontata da una nuova prospettiva. Il metodo *round-trip translation* (RTT) che consiste nel tradurre automaticamente un testo in un'altra lingua e poi di nuovo nella lingua di partenza, viene applicato in combinazione con BLEU, un metodo automatico per stimare la somiglianza dei testi tradizionalmente utilizzato nella valutazione della traduzione automatica. Calcoliamo la somiglianza tra il manoscritto e il prodotto della rispettiva RTT per ottenere una valutazione a priori della difficoltà di revisione e testiamo l'efficacia del nostro metodo. Insieme alla ricerca sperimentale vengono discusse alcune questioni teoriche riguardanti l'inglese come lingua franca e la pratica della revisione linguistica nel contesto dell'editoria accademica internazionale. I risultati degli esperimenti suggeriscono che il metodo RTT può essere utilizzato con successo per stimare la difficoltà di revisione. La varietà dei dati utilizzati negli esperimenti dimostra l'applicabilità del metodo al di là dei casi oggetto di studio. Concludiamo con una discussione delle limitazioni del lavoro e di possibili sviluppi futuri.

Parole-chiave: BLEU; inglese accademico; inglese come lingua franca; revisione; round-trip translation

1 Introduction

This study proposes an unconventional approach to a practical problem concerning the world of academic writing and publishing: the use of automatic round-trip translation to estimate the linguistic revision difficulty of texts to be published in English, used as a lingua franca.

The problem of pre-assessing the linguistic 'quality' of manuscripts and translations to be revised before publishing has been recognized as critical by various studies (e.g., Mossop 2001; Wang 2021). From a theoretical – and ethical – perspective, assessing the difficulty of linguistic revision needs to be conceived according to a pragmatic and functional vision of language, abandoning the constraints and prejudices deriving from prescriptivist ideologies. From a practical perspective, assessment is necessary to estimate costs and times of a revision. It has been noticed that the evaluation process can take substantial time and effort (Mossop 2001); hence, the possibility of providing this estimate automatically constitutes a new advantageous perspective for revisors, publishers and clients. Moreover, smoothing the process and costs of human revision could provide an alternative to, and thus prevent, an unconscious and unethical use of generative artificial intelligence (AI) and automatic writing assistants. The present work thus also aims to raise awareness by presenting both theoretical and practical reflections about revision practices, technology-aided academic writing and linguistic evaluation, which could be of interest to the new generations of revisors and translators.

By adopting a new point of view, which does not involve any prescriptive approach to correctness and evaluation, we investigate the possibility of evaluating the suitability of the English text by translating it. In the era of digital transformation, we do so by exploiting some of the most advanced technologies for automatic text processing, i.e., state-of-the-art machine translation (MT) systems through the method of round-trip translation (RTT) and automatic systems for text evaluation.

Round-trip translation is the process of translating a text from a source language into a target language, then translating the result back into the source language. This method has been previously used to evaluate automatic translation (Moon et al. 2020). Here it is used to produce similarity references to which the manuscripts under analysis have been compared through the use of two automatic scores for text similarity, which are traditionally employed in MT evaluation, namely BLEU (Papineni et al. 2002) and BERTScore (Devlin et al. 2018).

The dataset for our experiments is composed of texts that have been published and revised by a group of trainee translators at a publishing house based in Bologna (Italy), specializing in scientific and academic texts (Bologna University Press BUP).[1]

In this paper, an interdisciplinary review of literature about English as lingua franca in academic writing and linguistic revision is presented (Section 2). It is followed by a description of the method (Section 3) and a discussion about the results of the experiments that were conducted to test the feasibility of RTT for revision difficulty estimation (Section 4). In the closing section, some final remarks including ethical aspects and limitations of the study are provided.

2 Theoretical background

2.1 English as a lingua franca and academic writing

In the era of (post-)globalization linguae francae have developed as fundamental means for intercultural communication and knowledge sharing (Hall 2013). The massive spread of the English language in particular has raised many questions

[1] The revisions on which the experiments are based were done by a group of trainees from the Department of Interpreting and Translation of the University of Bologna, supervised by Ilaria Laurenza, a BUP revisor who had obtained her Master's degree at the same university department. Further information on revisors is given in the next note, while details about their workflow can be found in Wang (2021).

concerning its usage in all the different domains of human interaction (Jenkins 2013, 2015; Kachru 1992; Schneider 2012).

From a functionalist perspective (Ventola 1991), English as a lingua franca (ELF) can be defined as "any use of English among speakers of different first languages for whom English is the communicative medium of choice and often the only option" (Seidlhofer 2011: 7). Communicative behaviours, practices and interaction in ELF are continuously set and reset by users in the process of affirming their role as active speakers in different sociopolitical communicative contexts (McKinley and Rose 2018), this last aspect being crucial for a complete understanding of ELF characteristics and use (Lillis and Curry 2010).

This paper looks at ELF in connection with English for Academic Purposes (EAP) practices, defined by Hyland and Shaw as "language research and instruction that focuses on the communicative needs and practices of individuals working in academic contexts" (Hyland and Shaw 2016: 1).

Studies about ELF and EAP have frequently overlapped, both sharing a functionalist and social perspective on the use of English for transcultural communication (Bennett 2013; Hood 2016; Thompson and Diani 2015). These studies have problematized the traditional dichotomy between native and non-native speaker varieties, providing a decolonized perspective on writing and revision (Canagarajah 2002, 2023a). While the concept of *nativeness* is often – wrongly – perceived as directly related to proficiency in writing, research has generally acknowledged that skills in English academic writing must be acquired by both native and non-native writers (Bennett 2009; Römer 2009). Römer provides some valuable evidence in this direction, arguing that EAP is a rather specific case of non-native linguistic production, because "native speakers also have to learn […] academic writing," adding that "[t]he native academic writer does not seem to exist" (Römer 2009: 99).

In this respect, it should be noticed that the dichotomy between native and non-native speakers, far from representing the complex plurilingual and intersemiotic reality of communication (see, for example, 'translanguaging' theory by García and Wei [2014]), has conventionally been tied to monolingual ideologies which encourage cultural hierarchies and linguistic hegemony at different levels (Baird et al. 2014; Larsen-Freeman 2018; Seidlhofer 2011). Academic English is one clear example of a sociolect (Canagarajah 2023b) characterized by dominant native-speaker standards which have – at least until relatively recently – flattened the plurality of voices and standpoints in international scientific communities (Amano et al. 2023). Recently, however, scholars engaged with decolonizing practices in the use of English language(s) have proposed approaches and practical examples which subvert the standards of academic English by introducing unconventional language and mixing a plurality of codes and varieties in writing, maintaining their clarity, effectiveness and adequacy to the communicative context (Canagarajah 2023b). In line with such practices, our study proposes a methodology for estimating the need for linguistic revision based on objective factors related to functional communication, rather than ideological standards unveiling an imperial representation of native speakers as the only owners of English (Widdowson 1994), as will be discussed in the next section.

2.2 Language evaluation in the academic publishing industry and English linguistic revision

Against the background of research on academic text evaluation, especially focused on learners' assessment (Madnani and Cahill 2018; Yannakoudakis 2013), the studies on language standards and evaluation in scientific publication have mainly been developed by a niche of scholars who actively engage with the topic from the writer's perspective, as briefly mentioned at the end of Section 2.1.

In this landscape, studies on linguistic revision and the role of revisors are still underdeveloped. Consequently, this practice has been hardly defined and terminological ambiguity still characterizes the use of expressions such as 'revision', 'editing', and 'copy-editing' (Wang 2021).

For the purposes of this study, we broadly consider revision as the process of applying structural, lexical or grammatical changes to a text written and/or translated by somebody else to improve its adequacy and communicative effectiveness. It should also be noted that the borders between writing in ELF and translating into ELF are blurred, as many academic writers frequently engage with self-translation and automatic translation (O'Brien et al. 2018). Indeed, literature on linguistic revision is a collage of studies on writing, translation and editing for the publishing industry (Wang 2021), which overall show a functionalist and target-oriented approach. However, most studies focusing on the evaluation of English academic writing often underestimate, and sometimes completely forget, that such evaluation has a very practical aim: instructing the practice of linguistic revision.

When investigating revision difficulty estimation (i.e., linguistic evaluation for revision purposes), the following aspects need to be considered: the revisor's role in the publishing industry, types of revision and improvement techniques, and criteria for evaluation. The academic publishing companies' panorama is diverse, and the organization of the editing process changes from company to company, depending on the publisher's size and structural organization, on where the company is located and on the authority accorded to editorial boards (McGuigan and Russel 2008).

For what concerns the revisor's role in the publishing industry, it has been noted that revision is a heavily under-recognized practice. In the internal organization of the publishing houses it is very often a process to which not much attention nor a large budget are dedicated. In fact, in most cases, revisions are made by external experts, whose compensations are estimated based on the limited available budget (Mossop 2001). For example, Willey and Tanimoto (2013) provide insights on the working situation of English teachers as unexperienced revisors, who offer more convenient rates compared to those of professionals. Their study sheds light on the necessity for specialized training to overcome the *nativeness* bias on revisors' proficiency.[2]

The degree of revision (Mossop 2001), and, more generally, the choice of one approach or another do not depend only on the revisor's evaluation over the manuscript. Budget, time, requirements of the publishing house and target audience all play a role in the revisor's decision process. However, manuscript evaluation is still the starting point and essential step for any type of revision.

In recent years, the concept of readability has colonized the debate on text quality evaluation (Todirascu et al. 2016). In this respect, Mossop differentiates between *clarity* and *readability*, linking clarity to meaning and readability to the formal features of a text. But readability is a rather vague criterion, which leaves space for subjective and potentially biased evaluation.

Different views regarding ELF and EAP directly affect the approach to the revision task. Hartse and Kubota (2014) distinguish between error-based and variation-based approaches to revision, the latter including World Englishes, translingual and constructionist approaches. They discuss difficulties and issues of the revision process, noticing that "editing decisions reflect a complex interplay among native-speaker intuition, desire to avoid stigmatization, and conformity to standard varieties of English" (Hartse and Kubota 2014: 80).

The authors also note that there are very few academic journals which explicitly accept articles written in English varieties that diverge from Standard Native Speaker (NS) Englishes. They conclude that the "copyeditor is the final authority on appropriate language use" (Hartse and Kubota 2014: 79). Recently, some journals, especially in the field of linguistics and applied linguistics, such as the *Journal of English as a Lingua Franca*, published by DeGruyter, and the *ESP Journal* and the *Journal of Pragmatics*, published by Elsevier, do not strictly enforce native-like English proficiency and instead focus on clarity, coherence, and contribution to knowledge. These journals recognize the global nature of academic discourse and encourage submissions from non-native English speakers without requiring heavy linguistic editing, sometimes also asking for bilingual/plurilingual submissions. Still,

[2] In this respect, it may be worth mentioning that the team who provided the revision services for the texts under analysis was formed in the frame of an ongoing traineeship agreement between the Department of Interpreting and Translation at the University of Bologna and Bologna University Press. Moreover, students enrolled in Master's degree courses at the department can choose to participate in modules which also include specific training on the practice of revision.

linguistic revision by native speakers is required by many journals,[3] within and outside the field of linguistics, confirming that the issue of over-revision is still central and needs to be properly addressed (Billingham 2002; Tribble 2006). Revisors should critically think about the reasons for applying changes, to preserve "a greater variety in expression and more equity for writers", which in turn could result in "higher linguistic and ethical standard[s]" (Hartse and Kubota 2014: 81).

To fulfil this aim, nativeness should be replaced by some more purposeful concepts related to communicative success (Gosden 1995). For a successful communication, the language used in academia should account for "the regulating mechanisms of [academic] discourse community" (Gosden 1995: 38). In this sense, Mossop exploits the term *discourse expertise* to refer to both terminological precision and alignment with genre conventions (2001). However, while terminological correctness is easier to evaluate, genre conventions are highly variable and socially negotiated (Canagarajah 2022; Seidlhofer 2009). From a geopolitical point of view, Canagarajah affirms that "the representation of any research […] is considerably mediated by the rhetorical processes of writing and publishing" (Canagarajah 2002: 20). Adopting a similar view, Hyland and Shaw describe the results of globalization processes in terms of "inequality in the production and distribution of knowledge" (Hyland and Shaw 2016: 95).

In conclusion, the spread of English in academia and in the academic publishing industry means that most scientific texts are written, translated or self-translated in English by L2 users, for an L2 public. At the same time, a correlation between written English proficiency and research productivity has been noticed by Vasconcelos et al. (2007), making revision an essential practice for equal access to knowledge production.

Many scholars have proposed inclusive approaches to academic writing and revision. At the same time, some considerations on language standards are necessary because academic writing, for its own nature, needs to be comprehensible and appropriate to the meaning it aims to convey (as recognized also in studies which promote inclusive plural linguistic practices in academia, such as Balida et al. [2022] and Peters and Anderson [2021]). Therefore clarity, unambiguity and readability must be assured.

The complexity of the institutional and situational contexts points to the need for more structured linguistic services in the academic publishing industry which

[3] For example, the Taylor and Francis Group recommends that authors, particularly non-native English speakers, utilize academic editing services to ensure their manuscripts are "reviewed by a native speaker and the language polished before submission" (https://authorservices.taylorandfrancis.com/publishing-your-research/writing-your-paper/editing-services-improve-your-manuscript/, visited 1 February 2025)

should be aimed at creating bridges for scientific discussion rather than limits for publication. And while revision can serve as a means of empowerment and legitimation for authors whose work may otherwise be rejected simply because of language-related evaluation, inclusive views on languages, like translanguaging theories, should be encouraged within the academic publishing landscape as well as in translators and revisors training curricula, to enhance the use of non-standard varieties of English and other languages in academic communication (Canagarajah 2013; Schnell 2024).

2.3 Automatic text production evaluation, and automatic translation

In the context of automatic text processing, many studies have focused on quality evaluation of human linguistic productions. Probably the first approach to this topic as a stand-alone issue was related to the possibility of automatically assessing English learners' productions (Cotos 2009; McNamara and Graesser 2012; Yannakoudakis 2013). In general, learners' productions and academic writing in ELF share some features, but there is an important difference which makes it impossible to apply the same assessing methodologies on these texts, as highlighted in Section 2.2. In the publishing context, where the purpose of evaluation is very practical and target-oriented, the strongest concerns relate to linguistic standardization, simplification, and the imposition of native-like norms that may not align with authors' communicative objectives. Technology can undoubtedly streamline revision processes and improve objectivity, yet their reliance on fixed language models can introduce biases towards standardized output (Avner et al. 2016; Vanmassenhove et al. 2021), as also suggested by our last experiment (discussed in Section 4.4).

For this reason, the applications specifically meant for revision traditionally only support the revisors in the process; for example, text editors such as Libre Office and Microsoft Word are among the most common software tools employed by revisors (Billingham 2002; Mossop 2001). There are also plenty of online text editors, such as Google Documents and the innovative Collaborative Editing Tool for Non-Native Authors (CEPT), developed by Zhu et al. (2017).

More recently some applications have been created for automatic proof-reading. A well-known example is Grammarly, which has been identified as one of the most effective intelligent editing programs available (Max 2021).[4] Grammarly is a proprietary, cloud-based writing assistant designed for English text revision, employing a combination of grammar-based rules and machine learning techniques to provide

[4] https://www.grammarly.com (accessed in April 2022).

real-time feedback. The software's adaptive AI model refines its performance through user feedback, allowing it to improve error detection and style recommendations over time. Despite these advancements, AI-assisted proofreading tools are not neutral: they introduce implicit normative pressures, particularly in academic writing contexts, where evaluation metrics such as Grammarly's performance score may favor native-like textual conventions (Abu et al. 2024). Privacy concerns, potential biases towards simplification, and limited adaptability to ELF features are among the critical aspects that need further scrutiny.[5]

Automatic translation is another much-used technology in the publishing industry as it is in the academic world. Since its earliest stages (Shannon 1948; Weaver 1949), machine translation has changed a lot. State-of-the-art MT systems are based on neural networks, which make intensive use of deep learning techniques with bi-directional layers using end-to-end processes for decoding and encoding (Mikolov et al. 2010; Xue et al. 2014). The Encoder-Decoder (seq2seq) model was first used by Google in 2014; to date, it is one of the best performing neural network models for machine translation (Shah 2020). Thanks to this model, it is possible to produce outputs that are also very different from the inputs – a characteristic which is relevant to the use of MT proposed in the present work.

When it comes to MT quality evaluation (QE) two main approaches can be used: human evaluation, which is the most accurate but also expensive and time-consuming, and automatic evaluation (Rivera-Trigueros 2021). Progress in this research field is so rapid that it is unlikely that descriptions of MT architectures and performances will be found in textbooks. For the purposes of the present research, we have taken into account the *Proceedings of Machine Translation Summit* (Duh and Guzmán 2021) and the Intento report (2021), which were the latest available at the time the experiments were conducted. According to the report, ModernMT is one of the best-performing systems, with outstanding results in the education field.

Finally, it should be noted that the rapid advancements in large language models (LLMs) and the advent of generative AI have deeply affected text production tasks, including academic writing. AI tools such as ChatGPT have introduced new challenges for academic writing, particularly in the context of ELF and multilingual scholarship. While these AI-driven systems can assist with text structuring, coherence, and fluency, they also pose several risks in terms of over-standardization, loss of authorial agency, and potential ethical violations (Kocak 2024). Unlike the methodology proposed here, which is only meant to objectively evaluate the difficulty, hence time and effort needed, for human revision, generative AI systems work as fully automated writing assistants and their use in academic writing, although very

[5] https://support.grammarly.com/ (accessed in April 2022). In this study, we incorporate Grammarly's scoring system to compare the output of our method to already-existing ones.

recent, has already been attested and discussed (e.g., Bhatia 2023; Lin 2024; Zohery 2023). A comprehensive account of the issues related to the use of ChatGPT falls outside the scope of the present work. But given the centrality of such technologies in the current debate about text production, authorship, authenticity, scientific quality and linguistic standardization, a clarification is in order to highlight the differences between the methodology we propose and the assistence provided by generative AI.

First and foremost, ChatGPT and similar AI tools lack critical reasoning abilities and cannot engage in genuine scholarly argumentation (Bender et al. 2021). These models rely on statistical pattern recognition rather than an understanding of epistemic frameworks, research methodologies, or domain-specific knowledge. As a result, AI-generated text may contain plausible-sounding but incorrect, biased, or unverifiable information (Dale 2023). The problem of "hallucinated references", where AI fabricates non-existent citations, is particularly concerning for academic integrity (Thorp 2023). This issue raises serious ethical implications, as reliance on AI-generated content without rigorous verification could contribute to misinformation in scholarly discourse.

Furthermore, unregulated AI use in academic publishing risks compromising research authenticity, as AI-generated texts may be insufficiently distinct from existing sources, leading to concerns about self-plagiarism or unintentional content recycling (van Dis et al. 2023).

From an ELF perspective, there is also the danger that AI-assisted writing may result in linguistic homogenization, where the diversity of academic English varieties is eroded in favor of standardized expressions. This trend could diminish the agency of non-native scholars, reinforcing the idea that academic credibility is linked to native-like fluency rather than intellectual contribution (Myers 2023). AI-generated texts tend to produce outputs that align with monolithic, standard English conventions, potentially marginalizing diverse linguistic expressions and reducing the visibility of non-native academic voices. As AI models are trained predominantly on existing academic and internet corpora, they may favor dominant discursive patterns while overlooking alternative rhetorical structures commonly found in multilingual scholarship (Agarwal et al. 2024; Liang et al. 2023).

Potentially, a cautious use of such technolgies – which requires a thorough understanding of their functioning and risks – could also be beneficial but the above-mentioned issues and recent studies like Xiao et al. (2025) prevent us from considering generative AI systems as safe unbiased tools for linguistic revision.

Abolishing the use of generative AI in the academic world sounds like a paradoxical proposal, but alternative paths can be explored to promote critical use of AI. Following these considerations, the methodology described below does not seek to substitute human revisors, but to provide an output which can help them understand and predict the amount of effort and time necessary to enhance communicative effectiveness of academic texts, as will be clarified in the discussion of the results (Section 4).

2.3.1 Round-trip translation (RTT)

The possibility of using MT systems for academic writing purposes has become a reality since neural networks have started to produce outputs increasingly target-oriented. Such considerations were made by Groves and Mundt in a (2015) study that employed Google Translate – which still featured, at that time, a statistical architecture. Despite the substantial difference between statistical and state-of-the-art neural network MTs (Kalchbrenner and Blunsom 2013), Groves and Mundt's article demonstrates that already in 2014 Google Translator was able to produce texts which would meet – though only barely – international university standards. The study also pioneered research on machine translation applications for EAP. While collocating our investigation in this broader context, instead of focusing on automatic translation practices, we exploit MT systems to evaluate academic writing through the method of round-trip translation.

RTT is the process of translating a text from a source language into a target language (forward translation, FT), then translating the result back into the source language (back translation, BT), using machine translation (MT) software. We take inspiration from Moon et al. (2020), who propose this method for the evaluation of MT quality.

2.3.2 BLEU and BERTScore

Most traditional automatic methods for measuring text similarity treat texts as collections of words and use information about word frequencies to represent texts as vectors (Huang et al. 2011). Once two or more texts have been transformed into vectors, they can be compared with various methods that allow for text similarity estimation (Salton 1971).

In contrast to these first automatic techniques proposed for measuring text similarity, more recent methods are based on semantics. They expand semantically similar terms in the traditional word frequency vector and further increase the size of the text representation vector. Semantics-based methods estimate the similarity between texts accounting for semantic relationships such as synonymy and redundancy, but may not well reflect the strict similarity between two texts when it comes to formal differences. The two methods used in the present work – namely, BiLingual Evaluation Understudy (BLEU) and Bidirectional Encoder Representations from Transformers (BERT), constitute respectively an example of a well-established traditional method and an example of a newer semantic embedding method.

The BLEU method was proposed by Papineni et al. (2002) to evaluate MT systems. In the segment comparison method, individual words are used as the base unit for the comparison between candidate and reference. The scores obtained for the individual segments are then averaged over the entire text to obtain an overall estimate of the candidate's quality. The corpus score function is also available, which weighs each score before calculating the average. BLEU's output is a number between 0 and 1. A value close to 1 indicates high similarity while a value close to 0 indicates low similarity between the candidate and the reference.

BERTScore is "a language generation evaluation metric based on pretrained BERT contextual embeddings" (Zhang et al. 2020). BERT is a machine learning technique for natural language processing developed by Google in 2018 (Devlin et al. 2018). BERTScore "computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings" (Zhang et al. 2020: 1). Since it relies on cosine similarity, BERTScore is expressed with a number ranging between –1 and 1, with values close to 1 indicating high similarity.

Rescale functions are available for both BLEU and BERTScore. For example, BERTScore's rescale with baseline function, which was used in part of our experiments, allows for a rescale of the score within a natural range, thus making it comparable to the other scores.

3 Overview of the method and experimental setting

In order to examine the possibility of using round-trip translations as similarity references in the evaluation of academic writing revision difficulty, we made use of three datasets composed of parallel manuscripts and revisions from the BUP database.

The first dataset, "IE", includes manuscripts and revisions from the volume *The Italian economy after COVID-19: Short-term costs and long-term adjustment* (Bellettini and Goldstein 2020), which was published by BUP in 2020 after being revised by a group of revisors and trainees. This dataset was used for preliminary experiments which helped us define the best text similarity score for our application. The second dataset, called "CIHA", features texts on the topic of history of art (Faietti and Wolf 2021) and was used in the core experiments of this study. Besides the manuscripts and revisions, this dataset also features the revisors' evaluation and comments, which were used to categorize the texts. The revision task performed by the trainee revisors at BUP embraced a critical approach to revision nourished by theoretical seminars, university courses and discussions with BUP editors as well as academic staff at the Department of Interpreting and Translation. The revisors spoke different first languages, they could not rely on any native-speaker intuition; their decisions were primarily oriented towards terminological accuracy, clarity and linguistic adequacy, checked based on a plurality of corpora representing several English varieties. The head revisor was then asked to divide the texts into three sets according to the difficulty experienced during the revision process:

- **Set 1** includes texts which only needed proofreading and light editing. Manuscripts in this group were judged as overall well-structured, featuring an appropriate use of academic English.
- **Set 2** is composed of texts for which the revision required a fairly long time but no particular effort for the revisor, with many grammar errors and few structural issues.
- **Set 3** features texts for which the revision required a long time and a high mental effort; the manuscripts were grammatically incorrect, syntactically ambiguous, and presented inadequate use of terminology.

We refer to the categorization 1–3 as subjective qualitative revision difficulty index (IRev_S), whereby quality was estimated by the human revisor based on their own experience. Exploiting the *track changes* function in the revised files, a second parameter was obtained, named subjective quantitative revision difficulty index (IRev_Q). The average between IRev_Q and IRev_S provides IRev_H. This represents the reference parameter against which the automatic scores have been evaluated, which, as can be noticed, is not based on any a priori standards or idea of linguistic quality, but rather on the time and efforts employed by the revisors to improve the text readability, clarity and accuracy.

On IE and CIHA the first steps of our experimentation were performed as follows:

1. Data were converted into .txt files with UTF-8 encoding.
2. Manuscripts and respective revisions were segmented and aligned.
3. Round-trip translations of both the manuscripts and the revisions using ModernMT with English and Spanish as language pair were produced.[6]

After round-trip translation, IE and CIHA datasets contain five sets of texts: ‘Raw’ (the manuscripts), ‘Raw_RTT’ (the round-trip translations of the manuscripts), ‘Revisions’ (the revised versions of the manuscripts), ‘Revisions_RTT’ (their round-trip

[6] We chose English and Spanish since we know from the Intento report that ModernMT for academic/educational domain produces high-quality translations between English-Spanish and Spanish-English (Intento Inc 2021).

translations), and 'Revisions with comments' (where the .docx files with comments and changes made by the revisors are stored).

Then, we first ran some preliminary experiments on IE dataset to decide whether BLEU or BERTScore should be used to calculate automatic text similarity. BLEU has proven to be the best method to measure text similarity for the purposes of this study, because it highlights even small differences between the original files and their RTTs. Plus, it is time-saving and convenient in terms of computational power.

Secondly, to explore the potentials of automatic round-trip translation for revision difficulty estimation, we used the CIHA dataset and calculated the BLEU scores of different pairs of text:

a. manuscripts and their revisions;
b. manuscripts and their round-trip translations;
c. revisions and their round-trip translations;
d. revisions and manuscripts' round-trip translations.

Afterwards, the scores were interpreted through the cross-comparison of the scores deriving from the experiment, the human evaluation indexes, and Grammarly scores (used as an external automatic evaluation reference). It should be noted that, besides IRev_H, some more indexes were developed during the experiments to allow for a clearer visual comparison of the results, namely IRev_A, an automatic index of revision difficulty deriving from the automatic comparison of the manuscripts and revisions, and IRev_F, a final quality index which accounts for both IRev_A and IRev_H.

The following starting hypotheses were formulated with the aim of demonstrating the feasibility of our method:

1. The higher the similarity score between a manuscript and its revision, the higher the quality of the manuscript – if this is true, it means that BLEU is a good indicator of the amount and type of changes in a revised document compared to its first version.
2. The similarity score between a revision and its round-trip translation is higher than the score between the corresponding manuscript and its round-trip translation, in cases where substantial changes are required – if this is true, it means that round-trip translation quality is a predictor of linguistic quality and (inversely proportional to) difficulty of revision.

Finally, the results of these comparisons led to development of quality thresholds which were tested on our third dataset. This dataset contains papers from the journal *New Medit*, which were published without revision due to time and budget limitations. It was used to present a first application of our method on unrevised data.

Moreover, a further hypothesis linked to the possibility of using RTT as a starting point for revision has been tested on the CIHA dataset:

3. The similarity between the manuscript round-trip translations and their human revision is higher than the similarity between the manuscript and the revision – if this is true it means that the round-trip translation improves the manuscript similarly to a revision.

4 Results and discussion

4.1 First sets of experiments: BLEU scores between manuscripts and revisions compared to the human evaluation of revision difficulty

The first experiments were aimed at investigating hypothesis 1, i.e., that the higher the similarity score between a manuscript and its revision, the easier the revision.

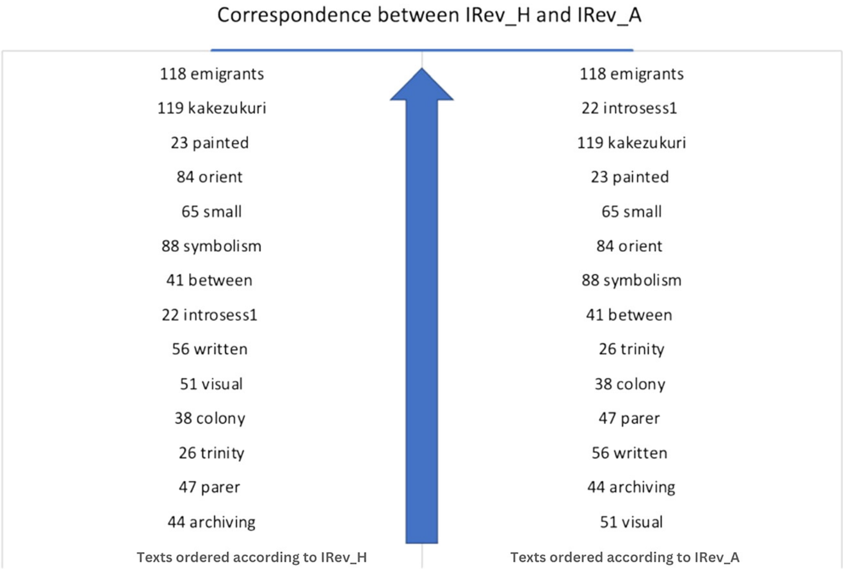
We computed BLEU scores between the manuscripts and their revisions. The scores were inverted (using the formula 1 – BLEU score) and normalized on a scale from 0 to 3 to be comparable with those deriving from the human evaluation; the resulting values are called Automatic Index of Revision Difficulty (IRev_A). The results of this experiment are shown in Table 1.

Table 1: BLEU score between raw texts and their revision. In the last column, the IRev_H is reported.

Text ID	Raw-Rev_BLEU	IRev_A	IRev_H
51 visual	0.980	0.176	0.950
44 archiving	0.872	0.344	0.759
56 written	0.961	0.600	1.414
47 parer	0.932	0.935	0.769
38 colony	0.888	0.988	0.881
26 trinity	0.894	1.129	0.857
41 between	0.742	1.588	1.838
88 symbolism	0.685	2.038	2.285
84 orient	0.769	2.038	2.595
65 small	0.820	2.276	2.440
23 painted	0.728	2.400	2.608
119 kakezukuri	0.702	2.629	2.629
22 introsess1	0.756	2.152	1.816
118 emigrants	0.660	3.000	3.000

In line with the preliminary experiments, these results demonstrate that BLEU can be employed to measure the similarity between a manuscript – the candidate – and its revision – the reference. Our first hypothesis is thus generally confirmed, although the classifications provided by IRev_A and IRev_H do not overlap perfectly, as can be noticed in Figure 1.

It should be noted that, in personal communication, the revisor has reviewed the task of evaluating texts according to a holistic judgement as "very difficult". In particular, she noticed that, while texts which needed major revision interventions could be easily separated from texts which did not, it was particularly hard to assign texts to the intermediate group. It could be observed that the same difficulty is reflected in the automatic comparison: BLEU scores are more accurate with clear-cut cases. A final index of revision difficulty (IRev_F) was derived from the average between the human and the automatic indexes, which is used to interpret the results of the subsequent experiments.



Correspondence between IRev_H and IRev_A

Texts ordered according to IRev_H	Texts ordered according to IRev_A
118 emigrants	118 emigrants
119 kakezukuri	22 introsess1
23 painted	119 kakezukuri
84 orient	23 painted
65 small	65 small
88 symbolism	84 orient
41 between	88 symbolism
22 introsess1	41 between
56 written	26 trinity
51 visual	38 colony
38 colony	47 parer
26 trinity	56 written
47 parer	44 archiving
44 archiving	51 visual

Figure 1: Comparison between texts ordered according to IRev_H and texts ordered according to IRev_A.

4.2 Second set of experiments: BLEU scores between manuscripts and their RTT as indicators of revision difficulty

The second set of experiments aims to verify hypothesis 2, namely whether the similarity between a manuscript and its round-trip translation is indicative of the difficulty – and, therefore, costs and time – of its revision. In order to test this possibility, the scores between manuscripts and their round-trip translation were compared to the scores between the respective revisions and their round-trip translations: Rev_RevRTT is expected to be higher than Raw_RawRTT in cases of substantial revision difficulty.

More precisely, we calculated the BLEU scores between the revised texts and their RTTs. Then, we compared these scores to the scores between the manuscripts and their RTTs. In order to make the scores comparable across the dataset, we calculated the difference between Rev_RevRTT and Raw_RawRTT. This difference has been inverted and rescaled on a range 0–3 to obtain Idx_Dis. Table 2 shows the results of the experiment, and the human evaluation values for comparison in the last column (IRev_H).

If our hypothesis is correct, the discrepancy index (Idx_Dis) is a good indicator of revision difficulty (Irev_H). Despite some differences, Figure 2 shows that the discrepancy index has similar values to the benchmark indices.

Table 2: BLEU scores of Raw_RawRTT and Rev_RevRTT.

Text ID	Raw_RawRTT	Rev_RevRTT	Rev_RevRTT – Raw_RawRTT	Idx_Dis	IRev_H
47 parer	0.777	0.733	−0.044	0.140	0.769
56 written	0.668	0.665	−0.003	1.101	1.414
51 visual	0.617	0.615	−0.002	1.125	0.950
44 archiving	0.662	0.661	−0.001	1.148	0.759
26 trinit	0.649	0.650	0.001	1.195	0.857
88 symbolism	0.646	0.658	0.012	1.453	2.285
38 colony	0.756	0.771	0.015	1.523	0.881
23 painted	0.685	0.705	0.020	1.640	2.608
41 between	0.668	0.689	0.021	1.664	1.838
65 small	0.592	0.617	0.025	1.757	2.440
119 kakezukuri	0.638	0.674	0.036	2.015	2.629
22 introsess1	0.648	0.687	0.039	2.085	1.816
84 orient	0.585	0.626	0.041	2.132	2.595
118 emigrants	0.588	0.666	0.078	3.000	3.000

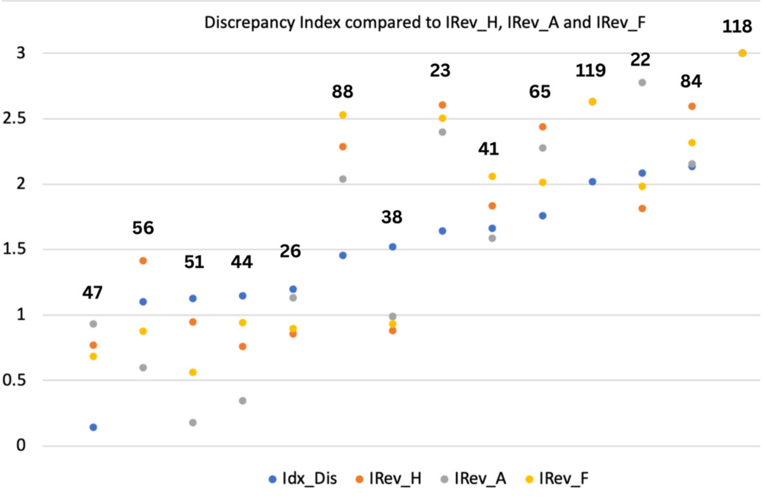


Figure 2: Comparison of discrepancy index (Idx_Dis), human revision, difficulty index (IRev_H), automatic revision difficulty index (IRev_A) and final revision difficulty index (IRev_F).

However, a closer analysis shows that there are some cases in which our hypothesis is not confirmed, namely when the manuscripts need little revision. Therefore, an ad hoc experiment on Text 47 was performed to examine the responsiveness of our method, by generating sample texts with randomly deleted words and added errors. Our results, as represented in Figure 3, show that the similarity between raw texts and their RTTs systematically decreases as text quality degrades.

The results in Table 2 were then compared to Grammarly scores. The comparison, presented in Figure 4, shows an agreement between the results of the RTT-based method and the scores provided by Grammarly (the yellow bars in Figure 4). A detailed representation of three texts representing the three groups based on the revisor's evaluation is also shown.

This comparison with an external benchmark seems to confirm the validity of the RTT/BLEU method. Thereby, a third set of experiments on unrevised data was conducted.

4.3 Third set of experiments: thresholds and testing

At this point in our experimentation, BLEU values had to be kept unaltered for the identification of BLEU thresholds. Consequently, we inverted IRev_H and normalized

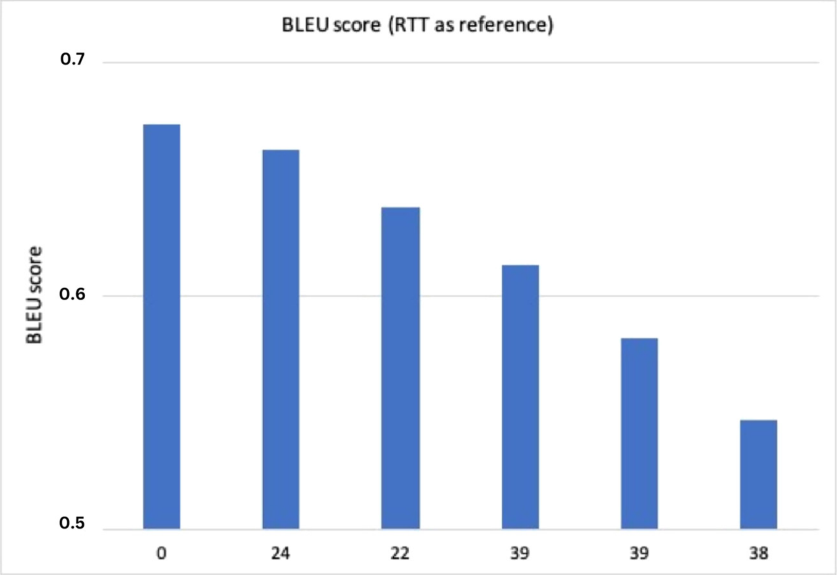


Figure 3: The BLEU scores for comparison of a sample reference text and candidates of decreasing quality. The number of randomly deleted words and/or deteriorating changes is reported on the *x* axis.

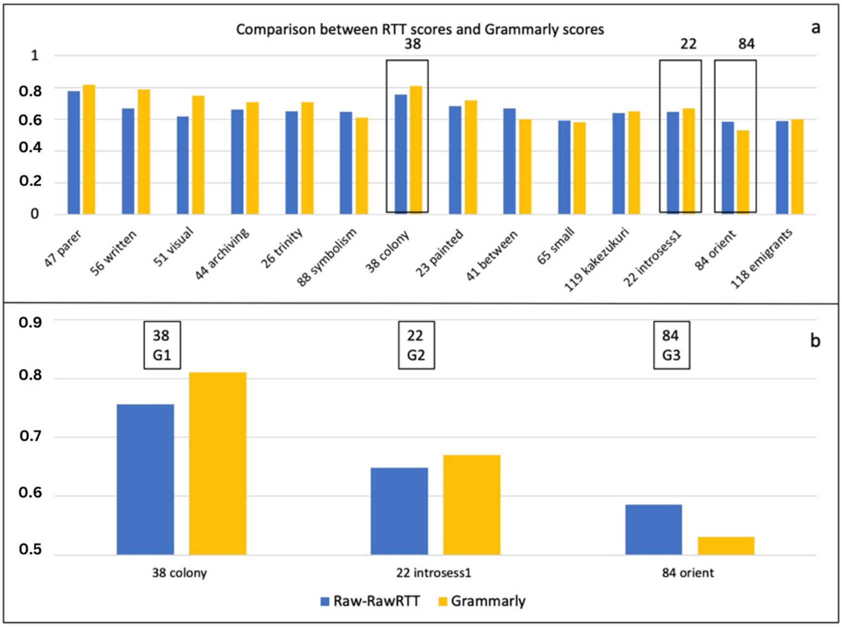


Figure 4: Comparison of BLEU and grammarly scores. (a) Comparison of the Raw-RawRTT score (blue bars) with the score provided by grammarly (yellow bars). (b) Details of the results of the experiments on three sample texts, compared with the grammarly score (yellow bars).

it on a scale from 0 to 1, to make it comparable to the BLEU scores. We refer to such values as IRev_E (index of revision ease). The correspondence between BLEU values and IRev_H/ IRev_F of the texts in CIHA allowed for a classification in three groups, based on observation of the visual clustering of the scores:

1. Texts with high revision difficulties characterized by a BLEU score ≤0.6.
2. Texts with low revision difficulty characterized by a BLEU score ≥0.75.
3. Texts with intermediate revision difficulty or with unpredictable revision difficulty for which the Raw versus Raw_RTT BLEU score is between 0.6 and 0.75.

As test set, three articles from the journal *New Medit* published by BUP were used. They are written in English as a lingua franca by non-native speakers, as confirmed by BUP editors, and they were not revised before publication. After the segmentation procedure, the round-trip translations of each text were generated in accordance with the method described in Section 3. Table 3 shows some sample segments and their round-trip translation.

Table 3: Sample segments extracted from the *New Medit* dataset.

Text ID	Raw	Raw_RTT
new_medit_1	This study drew on this gap and attempted to identify the UTPs that exist in the Cypriot food supply chain, assess their impact on the involved stakeholders and provide guidelines that will assist the transposition of EU relevant Directive to the national law	This study was based on this gap and attempted to identify UTPs that exist in the Cypriot food supply chain, assess their impact on the involved stakeholders and provide guidelines that will help the transposition of the relevant European Directive into national law
new_medit_2	Pragmatically speaking, we can say that ethnocentrism is the name given to the sense of belonging felt by any individual. More importantly, it explains the reason why a group accepts certain choices rather than others, which, in the case that concerns us here, refers to accepting certain buying behaviours rather than others	Pragmatically, we can say that ethnocentrism is the name given to the sense of belonging felt by any individual. More importantly, it explains why a group accepts certain choices over others, which in this case refers to the acceptance of certain purchasing behaviours over others
new_medit_3	This choice is in order to both take into account the nature of the decisions investigated and appropriately manage variables that can be complementary	This choice aims both to take into account the nature of the decisions studied and to manage appropriately the variables that may be complementary

This time, since we could not rely on posteriori automatic and – most importantly – human evaluations, the results were compared only to the Grammarly scores (normalized on a scale 0–1). The results are reported in Table 4.

As the table shows, with a value lower than 0.60, the new_medit_3 text belongs with the first group. On the other hand, new_medit_ 2 and new_medit_1 have BLEU scores of 0.717 and 0.645 respectively, thus falling into the intermediate range.

Moreover, Table 4 shows that BLEU on RTT and Grammarly scores are similar/correlated, while Figure 5 shows also the classification of the documents according to the groups defined above.

Table 4: BLEU scores of the test dataset (Raw_RawRTT [BLEU]) and the respective normalised grammarly scores.

ID Text	Raw_RawRTT(BLEU)	Grammarly
new_medit_1	0.645	0.720
new_medit_2	0.717	0.710
new_medit_3	0.528	0.510

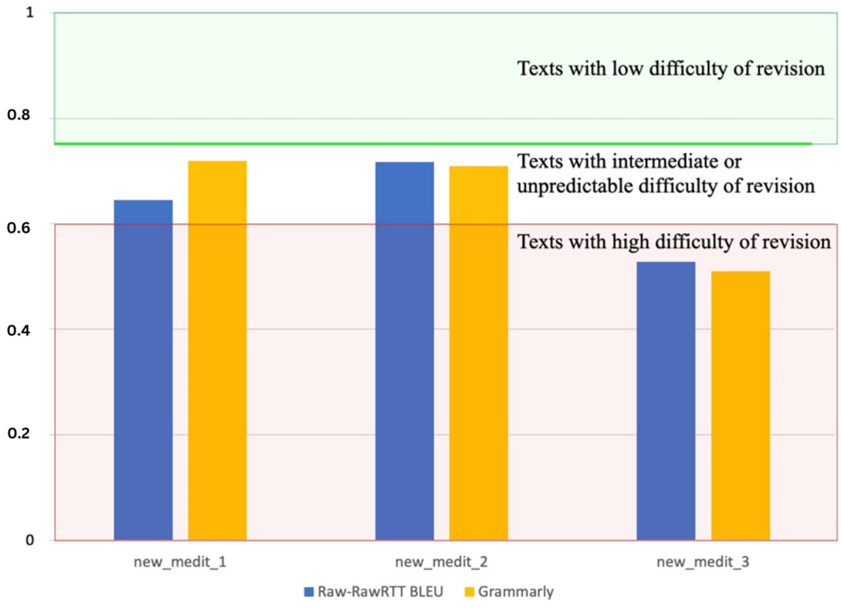


Figure 5: BLEU and grammarly scores for the *New Medit* test set.

These results suggest that the proposed thresholds for revision difficulty evaluation are consistent, although time and data availability limitations did not allow for further testing.

4.4 Experiments 4: evaluation of the RTT output

With the aim of better understanding MT behaviour in the context of round-trip translation, a fourth set of experiments was designed.

The similarity score between manuscript RTTs and revisions was computed and compared to the similarity between manuscripts and revisions. If the similarity between a manuscript's RTT and its revision is higher than the similarity between a manuscript and the corresponding revision, this would mean that RTT produces an output, thus it could be used as a starting point for revision. Figure 6 shows the results of this comparison.

The scores between the manuscripts (Raw) and their corresponding revisions (Rev) are systematically higher than the scores between the manuscripts round-trip translations (RawRTT) and the respective revisions (Rev). Thus, our third hypothesis has been disproved. These results indicate that the revisions by the expert revisor are

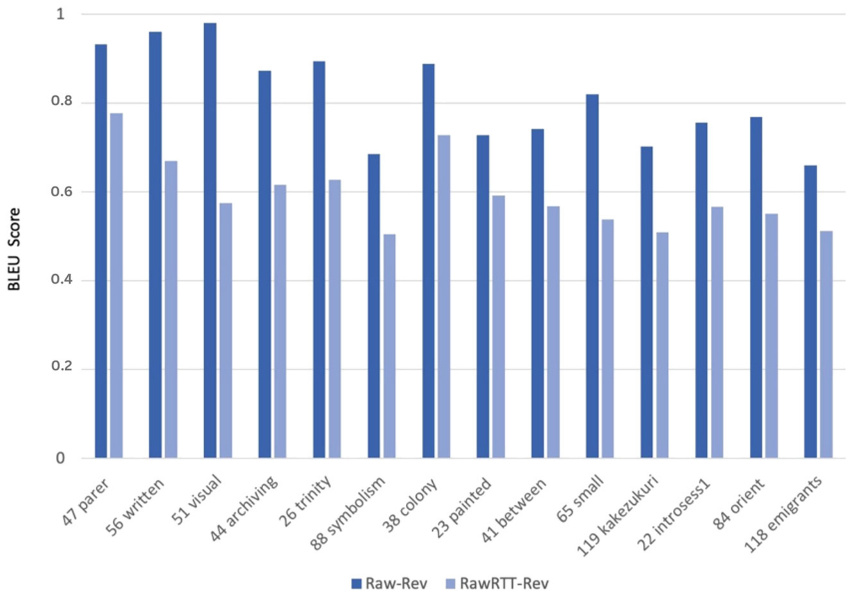


Figure 6: Comparison between Raw-Rev BLEU scores and RawRTT-Rev BLEU scores.

Table 5: Sample segments from text 38 (CIHA dataset).

Raw	RawRTT	Rev
The study starts with the Oath to Fernando VII, King of Spain, in 1809, and finishes examining some cultural events promoted by President Juan Rafael Mora Porras' government	The study begins with the oath to Ferdinand VII, King of Spain, in 1809, and ends by examining some cultural events promoted by the government of President Juan Rafael Mora Porras	The study starts with the Oath to Fernando VII, King of Spain, in 1809, and finishes by examining some cultural events promoted by President Juan Rafael Mora Porras' government
Then, they rang the church bells, many fireworks flew up and everyone shouted: ¡Long live the King don Fernando VII!	Then, the church bells ring, many fireworks fly and everyone shouts: Long live King Don Fernando VII!	Then, they rang the church bells, many fireworks flew up into the sky and everyone shouted: "Long live the King don Fernando VII!"

closely related to the source texts: the authors' structural and lexical choices are respected as much as possible. Table 4 provides an example extracted from text 38 of the CIHA dataset.

As can be noticed from Table 5, RTT produces random changes which are not necessary for the purposes of revision, and which – if applied by a human revisor – would be regarded as over-revision.

5 Final remarks and conclusion

In this work, we have addressed the problem of assessing and predicting the difficulty of revision that an expert revisor may encounter prior to the beginning of the revision process itself. This information, which normally requires a thorough analysis and decisions which can be deeply biased by the institutional context, is essential for planning the revision workflow and estimating its costs and duration. In line with the most recent approaches to linguistic issues, assistance for the solution of this problem has been found in automatic text processing. In this sense, we have tried to broaden the application of well-established techniques from the field of translation studies, with the hope that our methodology can provide less-biased indications and enhance critical thinking in revisors. Unlike automatic writing assistants, which tend to substitute writers and potentially reinforce language ideologies and Western dominant discourses, this use of technology complements human effort, thereby preserving human agency and differences.

Summing up, the RTT method for estimating the difficulty of revision has two main stages. The first is dedicated to the generation of automatic similarity references through round-trip translation, while the second involves the measurement of text similarity between the source text and the RTT using BLEU. The score which derives from this comparison represents an indicator of the a priori difficulty of revision: the higher this value, the lower the revision difficulty. Overall, the results of the experiments have confirmed our main hypothesis, according to which RTT can be a good indicator of revision difficulty. The proportional comparison of the automatic scores and the human evaluation suggests that BLEU scores are good indicators of the similarity between different versions of the same text. By comparing the difference between the similarity scores of Raw-RawRTT and Rev-RevRTT, an evaluating scale was obtained which resembles the one provided by the human revisor. Moreover, the comparison between the BLEU scores obtained through the method here tested and the Grammarly scores has validated the performance of our method.

Furthermore, the study triggered some reflections about the practice of revision as well as about machine translation. We have been able to confirm the suitability of the revision offered by the human revisors, while proposing a method which can provide assistance during the process of evaluation and revision. However, our results also show that the round-trip translation output cannot be used as a starting point for revision. In this respect, a qualitative analysis of the RTT outputs shows random lexical variations for both manuscripts and revisions regardless of their grammatical correctness and clarity.

One last aspect that we have tried to address relates to the ethical issues concerning the use of artificial intelligence in academia. In recent decades, automatic analysis methods based on computer systems have become increasingly widespread and established. Such a phenomenon is accompanied by many side-effects, for example language standardization, and privacy issues (Mayne 2021). It demands reflections on the changes that the fast technological progress has caused in human life and in the global society (Boden 2018; Madnani and Cahill 2018). In this respect, we believe that the debate around technology – especially since the advent of generative AI, should be restated in terms of users' awareness, technology-aided work practices and the narrowing of disparities. Our work eventually aims at exploring the possibilities offered by automatic methods to overcome practical and ideological issues related to language evaluation within the world of academic publishing. It does not suggest that automatic evaluation should entirely replace human critical thinking, but it can surely support the creation of objective, time-effective, human-led protocols which ensure clarity in academic writing while preserving cultural differences and linguistic variation.

Overall, why is it so hard to estimate revision difficulty? Because revising is ultimately the art of negotiation. Translation and revision are practices that share many common features: first and foremost, they both involve engagement with a text written by somebody else. With this parallelism in mind, the technological method here presented can be seen as a useful aid for clients to budget the costs of revision and for revisors to approach manuscripts in a more objective, functional and effective way.

Acknowledgments: We would like to thank the team at BUP, Ilaria Laurenza, who coordinated the traineeship, Luise and the other trainee reviewers who provided essential help and data for the study.

References

Abu Qub'a, Abdallah, Mohammed Nour Abu Guba & Shehdeh Fareh. 2024. Exploring the use of grammarly in assessing English academic writing. *Heliyon* 10(15). e34893.

Amano, Tatsuya, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White & Diogo Veríssimo. 2023. The manifold costs of being a non-native English speaker in science. *PLoS Biology* 21(7). e3002184.

Agarwal, Dhruv, Mor Naaman & Aditya Vashistha. 2024. AI suggestions homogenize writing toward western styles and diminish cultural nuances. arXiv:2409.11360. https://doi.org/10.48550/arXiv.2409.11360.

Avner, Ehud Alexander, Noam Ordan & Shuly Wintner. 2016. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities* 31(1). 30–54.

Baird, Robert, Will Baker & Mariko Kitazawa. 2014. The complexity of ELF. *Journal of English as a Lingua Franca* 3(1). 171–196.

Balida, Don Anton Robles, Romulo Alegre, May R. S. Lopez & Glennest J. D. Balida. 2022. Perspectives on Covid19 safety protocols among non-native English speaking teachers and students. *World Journal of English Language* 12(1). p419.

Bellettini, Giorgio & Andrea Goldstein (eds.). 2020. *The Italian economy after Covid-19. Short-term costs and long-term adjustments*. Bologna: Bologna University Press.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623. New York: Association for Computing Machinery.

Bennett, Karen. 2009. English academic style manuals: A survey. *Journal of English for Academic Purposes* 8(1). 43–54.

Bennett, Karen. 2013. English as a lingua franca in academia. *The Interpreter and Translator Trainer* 7(2). 169–193.

Bhatia, Pradeep.. 2023. ChatGPT for academic writing: A game changer or a disruptive tool? *Journal of Anaesthesiology Clinical Pharmacology* 39(1). 1–2.

Billingham, Jo. 2002. *Editing and revising text*. Oxford: Oxford University Press.

Boden, Margaret Ann. 2018. *Artificial intelligence: A very short introduction*. Oxford: Oxford University Press.

Canagarajah, Suresh. 2002. *A geopolitics of academic writing*. Pittsburg: University of Pittsburg Press.

Canagarajah, Suresh. 2013. *Translingual practice: Global Englishes and cosmopolitan relations*. New York: Routledge.

Canagarajah, Suresh. 2022. Language diversity in academic writing: Toward decolonizing scholarly publishing. *Journal of Multicultural Discourses* 17(2). https://doi.org/10.1080/17447143.2022.2063873.

Canagarajah, Suresh. 2023a. Decolonizing academic writing pedagogies for multilingual students. *Tesol Quarterly* 58(1). 280–306.

Canagarajah, Suresh. 2023b. *Resisting and enregistering norms in academic English*. Lecture at the University of Stockholm 15 June.

Cotos, Elena. 2009. Designing an intelligent discourse evaluation tool: Theoretical, empirical, and technological considerations. In *Developing and evaluating language learning materials, proceedings of the 6th annual TSLL conference*, 103–127. Ames: Iowa State University.

Dale, Robert. 2023. GPT-3: What's it good for? *Natural Language Engineering* 29(1). 1–15.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Toutanova Kristina. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (long and short papers)*. Minneapolis: Association for Computational Linguistics.

Duh, Kevin & Francisco Guzmán. 2021. *Proceedings of machine translation summit XVIII: Research track*. Virtual: Association for Machine Translation in the Americas. https://aclanthology.org/2021.mtsummit-research.0/ (accessed 24 February 2025).

Faietti, Marzia & Gerhard Wolf (eds.). 2021. *Motion: Transformation: 35th congress of the international committee of the history of arts. Florence, 1–6 September 2019. Congress Proceedings. 2 vols*. Bologna: Bologna University Press.

García, Ofelia & Li Wei. 2014. *Translanguaging. Language, bilingualism and education*. Berlin: Springer Link.

Gosden, Hugh. 1995. Success in research article writing and revision: A social-constructionist perspective. *English for Specific Purpose* 14(1). 37–57.

Groves, Michael & Klaus Mundt. 2015. Friend or foe? Google translate in language for academic purposes. *English for Specific Purposes* 37. 112–121.

Hall, Christopher J. 2013. Cognitive contributions to plurilithic views of English and other languages. *Applied Linguistics* 34(2). 211–231.

Hartse, Joel Heng & Ryuko Kubota. 2014. Pluralizing English? Variation in high-stakes academic texts and challenges of copyediting. *Journal of Second Language Writing* 24. 71–82.

Hood, Susan. 2016. Systemic functional linguistics and EAP. In Ken Hyland & Philip Shaw (eds.), *The Routledge handbook of English for academic purposes*, 193–205. London: Routledge.

Huang, Cheng-Hui, Jian Yin & Hou Fang. 2011. A text similarity measurement combining word semantic information with TF-IDF method. *Jisuanji Xuebao* 34(5). 856–864.

Hyland, Ken & Philip Shaw (eds.). 2016. *The Routledge handbook of English for academic purposes*. London: Routledge.

Intento Inc. 2021. *The state of machine translation*. Report.

Jenkins, Jennifer. 2013. The spread of English as a lingua franca. In Jennifer Jenkins (ed.), *English as a lingua franca in the international university*. London: Routledge.

Jenkins, Jennifer. 2015. Repositioning English and multilingualism in English as a lingua franca. *Englishes in Practice* 2(3). 49–85.

Kachru, Braj B. 1992. *The other tongue: English across cultures*. Champaign: University of Illinois Press.

Kalchbrenner, Nal & Phil Blunsom. 2013. Recurrent continuous translation models. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu & Steven Bethard (eds.), *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1700–1709. Seattle: Association for Computational Linguistics.

Kocak, Zafer. 2024. Publication ethics in the era of artificial intelligence. *Journal of Korean Medical Science* 39(33). e249.

Larsen-Freeman, Diane. 2018. Complexity and ELF. In Jennifer Jenkins, Will Baker & Martin Dewey (eds.), *The Routledge handbook on English as a lingua franca*, chapter 4. New York: Routledge.

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu & James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4(7). https://doi.org/10.1016/j.patter.2023.100779.

Lillis, Theresa & Mary Jane Curry. 2010. *Academic writing in a global context. The politics and practices of publishing in Englis*. London & New York: Routledge.

Lin, Zhicheng. 2024. Techniques for supercharging academic writing with generative AI. *Nature Biomedical Engineering* 1–6. https://doi.org/10.1038/s41551-024-01185-8.

Madnani, Nitin & Aoife Cahill. 2018. Automated scoring: Beyond natural language processing. In Emily M. Bender, Leon Derczynski & Pierre Isabelle (eds.), *Proceedings of the 27th international conference on computational linguistics*, 1099–1109. Santa Fe: Association for Computational Linguistics.

Max, Tucker. 2021. The best free & paid proofreading & editing software. *Scribe media* https://scribemedia.com/proofreading-editing-software/ (accessed January 2021).

Mayne, Dorothy. 2021. Revisiting Grammarly: An imperfect tool for final editing. In *Another word: From the writing center at university of wisconsin madison*. https://dept.writing.wisc.edu/blog/revisiting-grammarly/ (accessed January 2022).

McGuigan, Glenn S. & Robert D. Russel. 2008. The business of academic publishing: A strategic analysis of the academic journal publishing industry and its impact on the future of scholarly publishing. In *The Electronic Journal of Academic and Special Librarianship (1999–2008, volumes 1–10)*, 105. Athabasca: International Consortium for the Advancement of Academic Publication.

McKinley, Jim & Heath Rose. 2018. Conceptualizations of language errors, standards, norms and nativeness in English for research publication purposes: An analysis of journal submission guidelines. *Journal of Second Language Writing* 42. 1–11.

McNamara, Daniel S. & Arthur C. Graesser. 2012. Coh-Metrix: An automated tool for theo- retical and applied natural language processing. In Philip McCarthy & Chutima Boonthum-Denecke (eds.), *Applied natural language processing: Identification, investigation and resolution*, 188–205. Hershey: IGI Global.

Mikolov, Tomas, Martin Karafiát, Lukáš Burget, Jan Černocký & Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the 11th annual conference of the international speech communication association (INTERSPEECH 2010)*, 1045–1048. Makuhari, Chiba: International Speech Communication Association.

Moon, Jihyung, Hyunchang Cho & Eunjeong L. Park. 2020. Revisiting round-trip translation for quality estimation. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg & Mikel L. Forcadaeds (eds.), *Proceedings of the 22nd annual conference of the European association for machine translation*, 91–104. Lisbon: European Association for Machine Translation.

Mossop, Brian. 2001. *Revising and editing for translators, Translation practices explained*. Machester: St. Jerome Publishing.

Myers, Andrew. 2023. *AI-detectors biased against non-native English writers*. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers (accessed 24 February 2025).

O'Brien, Sharon, Michel Simard & Marie-Josée Goulet. 2018. Machine translation and self-post-editing for academic writing support: Quality explorations. In Joss Moorkens, Sheila Castilho, Federico Gaspari & Stephen Doherty (eds.), *Translation quality assessment: From principles to practice (machine translation: Technologies and applications 1)*. Berlin: Springer.

Papineni, Kishore, Salim Roukos, Ward Todd & Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak & Dekang Lin (eds.), *Proceedings of the 40th annual meeting of the association for computational linguistics*, 311–318. Philadelphia: Association for Computational Linguistics.

Peters, Bethany D. & Michael E. Anderson. 2021. Supporting non-native English speakers at the university. A survey of faculty & staff. *Journal of International Students* 11(1). 103–121.

Rivera-Trigueros, Irene. 2021. Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation* 56. 593–619.

Römer, Ute. 2009. English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies* 20(2). 89–100.

Salton, Gerard. 1971. *The SMART retrieval system – Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.

Schneider, Edgar W. 2012. *English around the world: An introduction*. Cambridge: Cambridge University Press.

Schnell, Bettina. 2024. Multilingual scholarly publishing: Exploring the perceptions, attitudes, and experiences of plurilingual scholars in foreign language publication. *Journal of Electronic Publishing* 27(1). https://doi.org/10.3998/jep.5416.

Seidlhofer, Barbara. 2009. Accommodation and the idiom principle in English as a lingua franca. *Intercultural Pragmatics* 6(2). 195–215.

Seidlhofer, Barbara. 2011. *Understanding English as a lingua franca*. Oxford: Oxford University Press.

Shah, Dhruvil. 2020. Machine translation with the seq2seq model: Different approaches. In *Towards data science*. https://towardsdatascience.com/machine-translation-with-the-seq2seq-model-different-approaches-f078081aaa37 (accessed February 2021).

Shannon, Claud Elwood. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423.

Thompson, Paul & Giuliana Diani (eds.). 2015. *English for academic purposes: Approaches and implications*. Newcastle Upon Tyne: Cambridge Scholars Publishing.

Thorp, Holden H. 2023. ChatGPT is fun, but not an author. *Science* 379(6630). 313.

Todirascu, Amalia, Thomas François, Delphine Bernhard, Núria Gala & Anne-Laure Ligozat. 2016. Are cohesive features relevant for text readability evaluation? In Yuji Matsumoto & Rashmi Prasad (eds.). *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 987–997. Osaka: The COLING 2016 Organizing Committee.

Tribble, Christopher. 2006. Written in, written out: Who sets the standards for academic writing? In Esther Usó-Juan & Alicia Martínez-Floran (eds.), *Current trends in the development and teaching of the four language skills (Studies on language acquisition [SOLA] 29)*, 447–472. New York: Mouton de Gruyter.

van Dis, Eva A. M., Bollen Johan, Willem Zuidema, Robert van Rooij & Claudi L. H. Bockting. 2023. ChatGPT: Five priorities for research. *Nature* 614(7947). 224–226.

Vanmassenhove, Eva, Dimitar Shterionov & Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Paola Merlo, Jorg Tiedemann & Reut Tsarfaty (eds.), *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume, 2203–2213*. Association for Computational Linguistics Online: https://aclanthology.org/2021.eacl-main/ (accessed 24 February 2025).

Vasconcelos, Sonia M. R., Martha M. Sorenson & Jaqueline Leta. 2007. Scientist-friendly policies for non-native English-speaking authors: Timely and welcome. *Brazilian Journal of Medical and Biological Research* 40(6). 743–747.

Ventola, Eija (ed.). 1991. *Functional and systemic linguistics (trends in linguistics. Studies and monographs)*, 55. Berlin & New York: Mouton de Gruyter.

Wang, Xiaoli. 2021. *Academic English editing in the era of COVID-19: A corpus-assisted case study*. Forlì: University of Bologna, Department of Translation and Interpreting MA Thesis.

Weaver, Warren. 1949. *Translation*. New York: Rockefeller Foundation.

Widdowson, Henry G. 1994. The ownership of English. *Tesol Quarterly* 28(2). 377–389.

Willey, Iain & Kazuko Tanimoto. 2013. “Convenience editors” as legitimate participants in the practice of scientific editing: An interview study. *Journal of English for Academic Purpose* 12(1). 23–32.

Xiao, Feiwen, Siyu Zhu & Xin Wen. 2025. Exploring the landscape of generative AI (ChatGPT)-powered writing instruction in English as a foreign language education: A scoping review. *ECNU Review of Education*. https://doi.org/10.1177/20965311241310881.

Xue, Shaofei, Ossama Abdel-Hamid, Hui Jiang, Lirong Dai & Qingfeng Liu. 2014. Fast adaptation of deep neural network based on discriminant codes for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(12). 1713–1725.

Yannakoudakis, Helen. 2013. Automated assessment of English-learner writing. In *Technical Report UCAM-CL-TR-842*. Cambridge: University of Cambridge, Computer Laboratory.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger & Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. arXiv:1904.09675.

Zhu, Yeshuang, Shichao Yue, Chun Yu & Yuanchun Shi. 2017. Cept: Collaborative editing tool for non-native authors. In *CSCW ’17: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 273–285. New York: Association for Computing Machinery.

Zohery, Medhat. 2023. ChatGPT in academic writing and publishing: A comprehensive guide. In *Artificial intelligence in academia, research and science: ChatGPT as a case study*, 10–61. São Paulo: Associação Brasileira de Divulgação Científica.