Journal of Audiovisual Translation

Human Agency in Live Subtitling Through Respeaking: Towards a Taxonomy of Effective Editing

Tomasz Korybski ✉
University of Surrey, Centre for Translation Studies

Elena Davitti ✉
University of Surrey, Centre for Translation Studies

Citation: Korybski, T., & Davitti, E. Human Agency in Live Subtitling through Respeaking: Towards a Taxonomy of Effective Editing. Journal of Audiovisual Translation, 7(2), 1–22. https://doi.org/10.47476/jat.v7i2.2024.302

Editor(s): N. Reviers, J. Neves & G. Vercauteren.

Received: November 17, 2023

Accepted: July 2, 2024

Published: December 19, 2024

Abstract

This paper examines the phenomenon of effective editions (EEs) as used by respeakers during live assignments. While the term "editing" conventionally refers to the refinement of written text, live spoken language editing has been recognised as a regular practice in the context of intra- and interlingual respeaking (Romero-Fresco & Pöchhacker, 2017). However, the existing definition of EEs could be expanded, and there are benefits to be reaped from a more comprehensive exploration of the range of phenomena encompassed under this umbrella term. This paper endeavours to fill the gap by scrutinising instances of EEs from an extensive database gathered within the framework of the ESRC-funded SMART project (ES/T002530/1, 2020–2023) on interlingual respeaking. Based on a bottom-up, empirical analysis, we propose a straightforward taxonomy of EEs, consisting of the main categories of condensation, re-expression, and compensation. Our analysis reveals the pervasive nature of EEs, which also emerge as significant predictors of respeakers' performance accuracy. The taxonomy we present is grounded in concrete examples and can facilitate a more equitable and pragmatic assessment of subtitle accuracy, but it also holds potential for refining (semi-)automated subtitle accuracy evaluation systems, which are currently at prototypical stages. Furthermore, the proposed taxonomy is relevant for respeaker training and/or upskilling where proficiency in effective editing will lead to enhanced performance. Given the nascent

✉ t.korybski@surrey.ac.uk, https://orcid.org/0000-0003-2353-0816
✉ e.davitti@surrey.ac.uk, https://orcid.org/0000-0002-7156-9275

status of research in this domain, the paper concludes by delineating prospective directions for further exploration of EEs.

Key words: respeaking, interlingual, intralingual, effective editions (EEs), subtitling, accuracy evaluation.

Introduction

In Translation Studies, including Audiovisual Translation Studies, the word "editing" commonly evokes associations with the process of machine translation post- or pre-editing. Pre-editing involves the pre-processing of texts before they undergo machine translation. This process generally includes rectifying errors in the source text, primarily related to grammar, punctuation, and spelling. In turn, post-editing involves trained linguists or translators reviewing and correcting machine translation output to eliminate semantic and linguistic errors. These practices tend to follow a chronological order (i.e., pre-editing precedes machine translation, and post-editing takes place after machine translation). However, in the context of *live* processing of spoken language input into textual output via speech recognition, such as in intra- and interlingual respeaking, there is an understudied aspect of synchronous effective editing performed by a human agent, which is the focal point of this paper. To explore it, we first need to outline the context of its application in real-life respeaking services.

Live subtitling is a key service to improve accessibility and inclusivity in various media and public settings, from TV broadcasts to live events such as conferences or lectures (Alonso-Bacigalupe & Romero-Fresco, 2023a). One human-centric way to provide live subtitles is through a workflow featuring a human respeaker working with speech recognition software. This workflow relies on efficient interaction between language professionals and AI-driven speech recognition technology and requires the deployment of complex skills, including concurrent listening and speaking, as well as monitoring and editing the textual output recognised and displayed by speech recognition software. This article aims to delve into one aspect of human-centric delivery of live subtitles through respeaking, namely the unique human ability to process and edit source content live in the context of respeaking for both intralingual and interlingual live subtitles.

Human respeakers possess unique skills that set them apart from automated transcription and machine-generated subtitles. Their ability to adjust to different accents, source speed, source text characteristics such as structure and lexical density, etc., leads them to shape and apply their live editing skills according to the needs at hand. By contrast, automated systems may struggle with real-time adjustments, leading to inaccuracies and misinterpretations (Bang et al., 2016). What drives human live editing is the fact that human respeakers possess a thorough understanding of the context and can accurately interpret colloquialisms, idioms, and cultural references that might be challenging for automated systems (Romero-Fresco, 2020) as well as filter-sensitive or inappropriate content, ensuring that the subtitles align with ethical and cultural norms (Bang et al., 2016). Not only that: respeakers intelligently use the features of live speech to simplify through reduction (Luyckx et al., 2010) and actively re-formulate content to ensure high quality and accuracy while considering the specific needs of the target audience at hand, making live subtitles more engaging and accessible (Romero-Fresco, 2020). Crucially, when live editing, respeakers also take into account the constraints of the output they are about to produce, i.e., the fact that it needs to be effectively processed through speech recognition software and fit the constraints of subtitles. Given the lack of one-to-one correspondence between a source and a target language, the need for effective editing may become

even more pronounced when language transfer is involved, as in interlingual respeaking. Interlingual respeakers require an extremely high level of linguistic proficiency in both source and target languages, as well as the ability to provide maximum accuracy in their renditions while maintaining cultural nuances and idiomatic expressions.

This paper broadens the concept of live editing to encompass more than just the reactive ability to promptly recognise and correct subtitle errors once processed and displayed on screen by speech recognition software. Our focus is on the proactive strategic behaviour implemented by respeakers at the stage of producing their "intermediary text" (Pöchhacker &Remael, 2019) to ensure the fitness-for-purpose of the target output. With current accuracy evaluation systems in this field primarily focusing on output errors, we have identified a gap in studies attempting to further define what "effective" editing in live contexts consists of.

1. Accuracy Evaluation in Respeaking

The importance of human-led live editing is acknowledged in live subtitle accuracy evaluation models like the NER (Romero-Fresco & Pérez, 2015) and NTR (Romero-Fresco & Pöchhacker, 2017). These models, departing from a word error-based assessment of accuracy (as in WER-type models, cf. Zue et al., 1990; Hunt, 1990), incorporate an editing dimension and recognise that the "live" nature of respeaking demands a more nuanced analysis of the target output.

The NER model, developed by Romero-Fresco (2011) and subsequently refined in Romero-Fresco and Pérez (2015), constitutes a recognised tool for gauging the accuracy of live subtitles generated through intralingual respeaking in the domains of media and live event broadcasts. The acronym NER encompasses the total word count in live subtitles (N), the count of edition errors (E), and the count of recognition errors (R). To ascertain the percentage of accurate content, the values of E and R are subtracted from N, and the result is divided by N. The NTR model, introduced by Romero-Fresco and Pöchhacker (2017), builds upon the NER model. It is specifically devised to assess interlingual respeaking. In the NTR model, edition errors are replaced by translation errors (T) to evaluate the accuracy of interlingual transfer.

Translation-related error categories under the NTR model encompass omissions, additions, and substitutions (content errors), as well as correctness and style (form errors). Notably, despite the varying nomenclature in each model, both the NER and the NTR employ the same categorisation of error severity, classifying errors as minor (0.25 penalty point), standard/major (0.5 penalty point), and serious/critical (1 penalty point). Importantly, distinguishing these models from WER-like systems, they accommodate Correct Editions (CEs, in the NER model) and Effective Editions (EEs, in the NTR model) within their analytical framework. Although both CEs and EEs are not included in the score calculation formulas, they play a role in the overall assessment of the output. In Romero-Fresco and Pérez (2015), *correct editions* include instances in which the respeaker's editing has not led to "a loss of information", whereby "the omission of redundancies and hesitations may be considered as

cases of correct editing and not as errors, as long as the coherence and cohesion of the original discourse are maintained." (Romero-Fresco & Pérez,2015, p. 6). In turn, in the context of interlingual respeaking and the NTR model, Romero-Fresco and Pöchhacker (2017, p. 159) define *effective editions* (EEs) as "deliberate deviations from the source text that do not involve a loss of information or that even enhance the communicative effectiveness of the subtitles". Consequently, the respeaker's *effective editing* skills are the abilities that allow them to provide "strategic reformulations" (Pöchhacker & Remael, 2019, p. 136) when confronted with content-related challenges in live assignments. These challenges can be of a different nature, including grammatical structure, vocabulary, redundancy, and stylistic deficiencies of the original.

Against the backdrop of these broad definitions, we posit that a deeper understanding of the human ability to effectively edit live spoken content is important. This paper provides an in-depth investigation of this phenomenon, seeking to define categories of effective editions and discussing the implications the taxonomy may have for both research and practice.

2. Strategic Reformulation in Live Settings: Simultaneous Interpreting and Respeaking

This section provides an overview of relevant research, focusing on strategic behaviours that guide live editing. In general, these strategic behaviours have been recognised in both interpreting and respeaking research (Pöchhacker & Remael, 2019). While this has been the object of extensive investigation in interpreting, research on respeaking is currently lacking a clear, data-driven taxonomy that systematizes the types of strategic reformulation practices performed by professionals. Attempting a categorisation of effective editing practices is crucial for understanding how respeakers manage to maintain and convey the original message within the constraints of a live setting and textual output. As respeaking and simultaneous interpreting have been found to be similar in many aspects, with a similar set of strategic reformulation behaviours applied by professionals when tackling and editing live content, we will start from the relevant notions explored in the existing literature.

Simultaneous interpreting refers to a practice where a human interpreter listens to spoken content and concurrently translates it into a target language, typically employing specialised equipment. Researchers have underscored the live aspect's close relationship with respeaking and the shared characteristics between the two practices (e.g., Davitti & Sandrelli, 2020; Eugeni, 2008; Robert & Remael, 2017; Romero-Fresco & Pöchhacker, 2017). Pöchhacker and Remael (2019) emphasise the interlingual and real-time aspects of these endeavours, along with the added challenge of editing. In the case of respeaking, there are added complexities resulting from the fact that respeakers must interact with the textual output displayed by speech recognition software, for example, monitoring recognition accuracy, correcting recognition errors, inserting punctuation and tags or finding ways to pre-empt specific problems from occurring, e.g. by reformulating ideas prior to feeding them into recognition software. Given the demanding immediacy of both simultaneous interpreting and respeaking, language professionals in both workflows must make swift decisions, often resolving

complex equivalence issues on the spot. To do this, they need specific skills and strategies, which have been approached from many angles but predominantly in Interpreting Studies. Researchers have offered some useful high-level approaches to these strategies. Pöchhacker (2003) categorises interpreting strategies into offline (happening before and after the interpreted event), and online strategies (relating to our focus in this paper, i.e., interpreter's performance during the event). In turn, in Riccardi's (2005) classification there are four high-level strategies pertinent to effective editing in respeaking, i.e., comprehension, production, overall and emergency strategies (exemplified by anticipation, generalisation, omission, and monitoring, as well as décalage or time lag management, respectively). The insight we offer in the present paper revolves around production strategies, as explained below.

Among the strategies identified under the broad umbrella term of "production", researchers have widely discussed segmentation and chunking (e.g. Jones, 1998), which consist of dividing the original utterance into several shorter segments or idea units that can be interpreted independently. Furthermore, transformation and reorganisation are frequently mentioned as production strategies where word order, sentence structure or even sentence order are changed in the target language (e.g. Bartłomiejczyk, 2006, Kalina, 1994; Kohn & Kalina, 1996; Moser-Mercer, 2000; Riccardi, 1996). In some cases, the output can be subject to deliberate paraphrasing, adaptation, or substituting achieved through adjusting word choices in the output based on the context, e.g., if a term is unknown in the target language or it is considered inappropriate in the target language's culture. Interpreters also apply approximation, i.e., finding the nearest possible solution by paraphrasing or using synonyms when the interpreter cannot access the "ideal" translation in time (e.g. Bartłomiejczyk, 2006; Kohn & Kalina, 1996). Output can also be extended when needed as a result of the interpreter's conscious expansion, addition and elaboration, as well as explicitation to refer to situations where a clearer and more precise target output is necessary. This can be achieved, for example, by using connectives to explicitate implicit or vague logic or by employing nouns for a corresponding pronoun in the source (e.g. Bartłomiejczyk, 2006; Gumul, 2017; Tang & Li, 2016). Compression, reduction, and condensation are at the other end of the spectrum of production strategies, leading to a more concise rendition in the target language by removing redundancy, often through omissions or "skipping" content (e.g. Bartłomiejczyk, 2006; Wang, 2012).

The selection of a particular strategy depends on the specific linguistic challenges, the context, and the language combination involved. Effective simultaneous interpretation demands quick thinking, adaptability, and a deep understanding of both source and target language. The same skill set is required from interlingual respeakers, so it is no wonder that they apply a similar repertoire of strategies. In a study looking at the performances of highly skilled linguists, Chmiel et al. (2017) zoomed in on paraphrasing skills in interpreters, respeakers, and bilinguals. The conclusion was that interpreters tended to excel in eliminating semantic redundancy and generating concise output, though the advantage was not consistently pronounced. Luyckx et al. (2010), in turn, focused on reduction practices in intralingual respeaking, highlighting that condensations and omissions appear to be deliberate processes influenced by external factors like source text speed and the availability of omissible respeaking units. This study also furnishes a compilation of source reduction strategies

that enable respeakers to preserve much of the original information despite substantial condensation (Luyckx et al., 2010, pp. 31–33).

In a rare study bridging the two professions, Sandrelli (2020) presents her classification of interpreter and respeaker interventions based on real-life data from a one-day symposium, concluding that the interventions of the language professionals involved (whether intentional or not) lead to three overarching categories of change in the output, specifically semantic transmission, reduction, and distortion. Semantic transmission pertains to occurrences wherein the semantic content of the source language (SL) idea unit is effectively conveyed by the target language (TL) idea unit. Reduction encompasses scenarios wherein certain information is omitted in the TL message, expressing only a portion of the SL content. Distortion denotes the factual alteration of semantic content, signifying instances in which the TL unit conveys a divergent idea from the original. Each of the categories is then subdivided into detailed types reflecting the relevant phenomena found in the product. In this paper, we build on the existing repertoire of strategies derived from interpreting and respeaking research, aiming to develop a simple and easy-to-apply taxonomy.

The strategy identification studies presented above form a foundation for our approach to defining and categorising EEs in respeaking. The following section presents the data and methodological approach that led us to our taxonomy of effective editing phenomena.

3. Dataset and Approach for Capturing and Analysing Effective Editions

This discussion is framed within the context of a major research project implemented at the University of Surrey, Centre for Translation Studies. SMART (*Shaping Multilingual Access Through Respeaking Technology*, Economic and Social Research Council UK, ES/T002530/1, 2020–2023) aimed to explore interlingual respeaking as a novel method for delivering real-time speech-to-text services across languages. The project analysed the performances of 51 language professionals with diverse backgrounds, encompassing consecutive and simultaneous interpreting, written translation, pre-recorded subtitling, and live subtitling. The study sought to investigate the competencies involved in interlingual respeaking, the accuracy achievable by language professionals with a minimum of 2,000 hours of practice in at least one of the disciplines, and strategies for enhancing upskilling based on empirical research insights.

The study employed a mixed-methods, experimental design to achieve its objectives, integrating qualitative and quantitative methods for data collection and triangulation. The analysis of interlingual respeaking performance accuracy and the associated self-reflective data are relevant to this paper. The experiment was centred around a customised 25-hour upskilling course in interlingual respeaking, conducted online over five weeks in a self-taught manner, prompted by the project's initiation during the pandemic. This course served the dual purpose of data collection for the study and providing participants with the same exposure to interlingual respeaking, in the form of an advanced introduction. Given the innovative nature of the practice and the scarcity of fully trained

professionals, the course was designed to accommodate language professionals from various backgrounds, each contributing unique skills to this emerging field (Wallinheimo et al., 2023).

As regards the language professionals' sample, there were eight males and 43 females. The mean age was 40.12 years (*SD* = 10.97 years), and the age range between 23 and 65. The participants joined the course from 11 countries (UK, Spain, Italy, France, Germany, Belgium, Australia, Argentina, New Zealand, USA, and Peru). After completing the 25-hour bespoke course, during which they were exposed to both intralingual and interlingual respeaking, participants were required to undergo some testing, namely respeak into their native languages (the language combinations were restricted to English to/from Italian, French, and/or Spanish). Each participant underwent six tests in total, encompassing three intralingual and three interlingual scenarios, designed to simulate three different conditions reflecting typical real-life scenarios in respeaking: speed (fast speakers), planned/unplanned delivery (partially improvised, partially prepared speech), and multiple speakers (interview scenario). All speeches were initially prepared in English and subsequently translated and adapted for fluency across the other languages. This approach ensured maximum comparability and enabled us to maintain consistent difficulty levels across all language pairs. All language pairs included English.

Participants were instructed to screencast their performances during each task. The SMART project methodology incorporated the use of retrospective Think Aloud Protocols (TAPs) to gather participants' self-analysis of their respeaking performance. This approach provided valuable insights from the professionals themselves, who recorded their reactions while reviewing screencasts of each performance.

The taxonomy presented in this paper stems from a data-driven, bottom-up analysis of the effective editions captured across SMART's interlingual dataset, comprising 153 performances lasting 15 to 20 minutes each, totalling over 2,500 recorded and analysed minutes. This dataset is a valuable source of interlingual respeaking data, with aligned sources and targets for each language pair and directionality analysed using the NTR model for accuracy evaluation (Davitti & Wallinheimo, 2024). To conduct this fine-grained analysis, we used a purpose-made assessment grid based on the NER score spreadsheet employed by Canadian media companies for intralingual respeaking evaluation (for more information, see Davitti & Sandrelli, 2020). These adapted grids facilitate source content segmentation, alignment of source and target idea units, and the scoring of translation (content- and form-related) and recognition errors. We also added a specific column for the qualitative and quantitative capturing and documentation of all instances of EEs.

For optimal consistency, each NTR sheet was assessed and reviewed by two evaluators per language pair competent in the relevant languages and trained by the research team in the use of both the NER and NTR models. The evaluators were language professionals with experience in different language-related practices, including respeaking, subtitling, translation and interpreting, as well as trainers and researchers/academics in these fields. Prior to the evaluation phase, they all underwent extensive training in both theoretical and hands-on data sessions to ensure a consistent approach to

the implementation of both models. Apart from capturing all errors to be penalised (omissions, additions, substitutions, style, form and recognition), all instances of EEs were also counted and commented on with due diligence as part of a double-evaluation approach. As explained earlier, EEs are not subject to weighting in either the NER and NTR models – they constitute additional information about the respeaker's performance and output characteristics. Each NTR file contains a sum of all its instances of EEs, which enabled us not only to track particular instances with their comments, but also to compare the total number of occurrences of EEs across scenarios and languages, both qualitatively and quantitatively.

After completing the NTR analyses of all performances and language directionalities, we identified as many as 6,175 instances of EEs. Following the guidelines in Field (2024), we also conducted statistical analyses in SPSS (version 29.0.1.0) to explore the relationship between EEs and NTR accuracy (both continuous variables). A linear regression model was used (without random effects or interactions). The result was significant $F(1, 49) = 5.36$, $p = .03$, indicating that the higher the number of EEs, the higher the NTR accuracy. The adjusted R2 indicated that the model explained 10% of the variance in NTR accuracy. There was a positive relationship between EEs (independent variable) and NTR accuracy (dependent variable) across all source speech type scenarios used ($t(50) = 110.06$, $\beta = .31$, $p = .03$). Additionally, a one-way ANOVA ($F(1, 49) = 4.71$, $p = .04$) revealed that high performers used EEs more ($M = 43.19$, $SE = 1.90$) than low performers ($M = 37.17$, $SE = 2.02$). These findings confirm the positive impact of EEs on performance, which underscored the need for additional qualitative investigation.

Based on these results, we selected a dataset of the highest NTR-achieving performances (ranging between 98.6% and 96.2%, representing around 20% of all performances). Data-driven, bottom-up analysis of the dataset showed a diverse yet consistent picture: it became apparent that the EEs from our data encompassed a diverse range of strategic reformulations executed by language professionals in their respeaking tasks. Multiple assess-talk-assess rounds modelled on the Delphi method (Kaplan et al., 1949) revealed recurrent patterns in live editing phenomena that we were able to group to form a foundation for the taxonomy.

Subsequently, we complemented our product-oriented observations with self-reflective comments from the TAPs. These contained rich information on the perceived challenges encountered by participants during their respeaking performances, providing insights into the effective editing strategies used and why. We mapped those onto our collection of preliminary categorised EEs to add more depth to the analysis. This set of findings was used as an auxiliary source in developing our taxonomy. Examples of relevant comments from the project's TAPs accompany our descriptions of taxonomy categories in section 4.

To test the categories also in the context of intralingual respeaking, we used some examples from a pilot project implemented in the years 2020–2022, MATRIC (*Machine Translation and Respeaking in Interlingual Communication* – for more details on the project see Korybski et al., 2022).

4. Developing the Effective Editions taxonomy

With the taxonomy, we intend to provide a systematic framework for categorising and understanding some key recurring moves captured across the instances of the EEs we collected. Notably, our goal is not to provide an exhaustive catalogue of all existing EEs, but rather to initiate a systematisation of this diverse set of phenomena based on a streamlined and easy-to-apply categorisation, open to expansion and refinement through future empirical studies.

Our taxonomy development was guided by a key observation. As we systematically examined EEs in the dataset, we noticed a certain correspondence between the negative error types penalised in the NTR (namely content-related errors associated with language transfer) and positive 'mirror reflections' moves identified in the EEs, as presented in Table 1.

Table 1

Correspondence of Error Types With EE Categories

EE CATEGORY	CORRESPONDING NTR ERROR TYPE and EXAMPLE	
CONDENSATION	→ Omissions	e.g. redundant ideas, phatic language, deictics, modifiers, adverbs, adjectives, connectives
RE-EXPRESSION		
Semantic level (content)	→ Additions → Substitutions	e.g. explicitation, specification e.g. implicitation, generalisation, adaptation
Lexical level (form) Structural level (form)	→ Style → Grammar	e.g. technical term expressed in simpler language e.g. syntactical restructuring
COMPENSATION		Making up for previously missed content across idea units (macro-level)

For example, the negative error type of omission is normally penalised in the NTR model as it captures situations where the respeaker omitted (more or less) significant content from the source, with (some) detriment to the message conveyed to the recipient of the subtitles. However, in interlingual respeaking, as in other live practices, particularly those involving language transfer like simultaneous interpreting, there can be instances where content omissions are positive. These omissions, while not resulting in the loss of conveyed information, contribute to an advantageous condensation of the target. This condensation may enhance the suitability of the resulting live captions, as condensation may lead to better readability of shorter but still meaningful live subtitles. Based on this observation and building on existing research on strategic behaviour in live communication practices, we found

that condensation could be a useful umbrella category for the different positive EEs sharing these characteristics.

We also identified another supra-category, re-expression. Based on examples from our data, re-expression includes three levels. A semantic level mirroring NTR content-related translation errors and exemplified by a successful application of substitutions and additions in EEs. For instance, added elements aiming to explicitate or further specify an element from the source speech, resulting in clearer output, or a substitution successfully generalising or adapting the target output, without loss of meaning. The second one is a lexical level, which mirrors form-related style errors, occurring, for instance, when a rare or technical term is expressed in simpler language for easier understanding. The third one is a structural level (restructuring), mirroring form-style grammar errors. At a micro level, this can be exemplified by syntactical interventions in successful EEs, for instance, when a rhetorical interrogative is changed into an affirmative statement conveying the same meaning. At a macro level, this can be exemplified by information reordering between adjacent idea units for enhanced readability and comprehension.

Finally, the third supra-category is compensation. This category does not mirror any NTR error type, but is frequently found in our dataset, where missing content is provided by the respeaker in a later fragment with no detriment to the overall message provided in the live captions.

We also differentiate between micro and macro dimensions of EEs. The former refers to EEs found within one idea unit, where the edition is self-contained and does not cross idea unit boundaries. In turn, the macro dimension, predominantly linked to the category of compensation, occurs when the EE taps into the local context of the fragment at hand and cuts across multiple segments (a minimum of two).

4.1. Condensation

Condensations appear to be a prominent category in our datasets. At the micro level, they occur when a source idea unit is compressed and expressed in a shorter form in the target idea unit without (any considerable) loss of information. In our data, condensation was implemented primarily through omissions of redundant information, phatic language, deictic expressions, or grammatical interventions. Condensations at the "macro" level occur when a target idea unit can be expressed more concisely by referring to the preceding idea unit or utilising elements of content from it, and when a target idea unit captures more than one source idea unit. This type of condensation is typically achieved using pronouns to replace names, or deictics such as "this", "that", "these", "those", "now", "then", "here".

Some participant comments from the TAPs confirm the existence of the category and reveal some of the reasons behind the use of condensation, which are mostly linked to the complex nature of respeaking:

Participant 116: "Throughout the video, I just skipped the interjections and confirmations to keep the pace".

Participant 123: "I managed to get the general gist of what they were saying, and therefore, as I was monitoring the speech, I ended up paraphrasing or summarising a lot".

Examples 1 and 2 below show two types of condensation sourced from the datasets. The source data transcriptions contain all the words that were, in fact, uttered by the original speaker. As a result, repetitions and redundancies are present in the source transcripts to demonstrate the full scope of the human intervention in the target.

Example 1 (intralingual respeaking)

Source: Despite the the dramatic and I would even say tragic events **we are just going to discuss in a minute** the first time we see each other after the after the Christmas break and **therefore** I **really** would like to wish all of you **all the citizens you represent and all the European Union** all the best in the New Year and happiest 2020

Target: Despite the dramatic and I would even say tragic events, it's the first time we have seen each other after Christmas and I would like to wish all of you all the best in the New Year and the happiest 2020.

Example 1 was significantly condensed by the respeaker. This resulted in shorter, more readable captions, with the message largely intact. Thanks to the respeaker's intelligent interpretation of the context, they were able to leave out a whole string of words that did not contribute important information in the context (e.g., the modifier "really", the reference to EU citizens who are indeed represented by the Members of the European Parliament), thus streamlining the entire process for themselves and for the recipients. Although some redundant elements, such as the repetitions resulting from hesitation or uncertainty, would have been cut out by currently available automatic speech recognition (ASR) tools, this depth of intervention could not be expected from an ASR solution.

Example 2 (interlingual respeaking, Spanish>English)

Source: **Entiendo que puede** ser algo confuso así que **quizá** debería explicarlo mejor

BT: *I understand that it can be somewhat confusing, so perhaps I should explain it better.*

Target: It might seem confusing, so I should clarify that.

The speaker's remark in Example 2 was made after an extensive explanation of the respeaking workflows in certain countries, which involve quite a few people performing different roles. This example features condensation on two levels: firstly, the rapport-building "I understand that" which is dropped in the target rendition. Secondly, "quizá / perhaps" is dropped, leaving the transfer of the sentence's modality to the modal verb "should". Arguably, these changes are justified: dropping the relational "I understand that" in this context does not impact the overall sentiment of the utterance and helps streamline the context for effective communication. Similarly, removing the softening

effect introduced by "perhaps" results in a more direct affirmative clause, which enhances the clarity and immediacy in conveying information. Overall, and in the wider context of this example, these modifications bring no detriment to the original's message and result in improved readability. Although there may be contexts where even a slight shift in modality will interfere with the message conveyed (e.g., a legal context), the respeaker's judgment of the situation (often supported by an assignment brief and preparation) should limit editing to the fragments of the source text that lend themselves well to such transformation.

The instances of condensation presented above show the uniquely human skill of editing based on split-second decisions and contextual judgment that result in well-readable output either in the same or a different language. Although today's large language models and text summarisation solutions cope well with text reduction (leading to condensation), at least two major areas of human superiority remain. Firstly, apart from redundant content, all content words in automatically summarised fragments of texts tend to stay in the output. Humans, in turn, are able to "filter out" non-essential words, including content words, based on the context. This results in succinct and easy-to-read captions. Secondly, human respeakers can switch their condensation practice on and off very dynamically within one assignment, while an automated solution would need to be prompted separately for selected fragments of the source speech to be able to mimic a human respeaker's behaviour. These two challenge areas, of course, come on top of all the existing challenges related to ASR (such as overlapping speech and background noise) providing input for any human-like editing operations in a cascaded system.

4.2. Re-Expression

Re-expression is about effectively using the lexico-semantic and structural possibilities afforded by the source content to produce successful renditions suitable for live captions. Re-expression can take the form of semantic, stylistic, and syntactic interventions that do not interfere with the message and provide a more readable output thanks to, for example, voice change, sentence splitting, sentence merging or additions or substitutions that are deemed positive and are, therefore, not penalised in the NTR evaluation.

Some participant comments from the TAPs confirm the existence of the category and reveal the rationale behind the use of re-expression, as in the quotes below, pertaining to structural and lexical intervention:

Participant 155: "I think I'm satisfied with a slight improvement and chunking".

Participant 189: "I think that was a quite good substitution, 'wrong temperature' instead of 'too hot or too cold', I'm quite impressed that it came out of my mouth".

Examples 3 and 4 below exemplify re-expression strategies implemented at a micro level through minor additions, positively affecting the target output. For instance, in Example 3, the addition of "When we speak of access" at the start of the rendition aligns perfectly with the tone and message of the previous segments, reinforcing the speaker's communicative intention and achieving an effective target output.

Example 3 (interlingual respeaking, French > English)

Source: Nous parlons donc d'accès pour tous.

BT: *So we are talking about access for all*

Target: **When we speak of access**, we speak of access for all.

Similarly, Example 4 shows an addition ("settoriale / sector-specific"), which is unnecessary in terms of content as the previous idea has been successfully conveyed in a more literal manner ("tema molto specifico / a very specific subject"). While additions are rare in the dataset, likely due to the cognitively challenging and real-time nature of the interlingual respeaking, effective additions like the one shown here can enhance target output clarity and idiomaticity.

Example 4 (interlingual respeaking, English > Italian)

Source: Let me tell you about a conference […] which was all about plants. It was a very specialised subject

Target: […] che aveva un **tema molto specifico, settoriale**

BT: *…which had a very specific subject, sector-specific*

Example 5 is an interesting instance of re-expression through substitution, resulting in an effective formulation in the target language. The source is the same as in Example 2, but in English, as this participant worked from English into French (see section 3 for how the source speeches were designed).

Example 5 (interlingual respeaking, English > French)

Source: I know it can be perhaps a bit confusing**, so maybe I should explain more**.

Target: Je sais que cela peut être assez confus, donc je **vais vous expliquer en quoi cela consiste.**

BT: *I know that this can be quite confusing, so I will explain to you what it consists of.*

The concise "I should explain more" is unpacked in French into "I will explain (to you) what is consists of", thus showing an instance of semantic re-expression through two types of substitutions. "*More*" rendered as "en quoi cela consiste" (what it consists of) enhances the clarity of the target, while maintaining a fluent formulation in French. Stylistically, the substitution of the modal verb "should" into a future indicative tense does not modify the final meaning but explicitates the speaker's intention to further elaborate on the content.

Example 6 shows an instance of stylistic re-expression, where two concepts ("respeakers" and "working interlingually") are reworked through paraphrasing. Based on the participants' TAP comments, this was done to prevent issues with recognition.

Example 6 (interlingual respeaking English > French)

Source: So the **respeakers** would be working **interlingually**,

Target: C'est-à-dire que les **personnes qui font le respeaking** parlent **dans une autre langue**.

BT: *That's to say that the people who are doing the respeaking speak in another language.*

Example 7 (interlingual respeaking Spanish > English)

Source: Entonces, ¿significa esto que las personas que se dedican al rehablado intralingüístico pues que no tienen experiencia como intérpretes pueden dedicarse al rehablado interlingüístico?
Yo diría que sí, absolutamente

BT: *So, does this mean that people who perform intralingual respeaking, since they do not have experience as interpreters, can still perform interlingual respeaking?*
I would say yes, absolutely

Target: So this means that people who are intralingual respeakers who are not interpreters can work as interlingual respeakers if they work hard**.**

In Example 7, we can see a structural re-expression: the respeaker changed the source's rhetorical interrogative into an affirmative (the shift was possible, as the answer was provided by the speaker in the next sentence). Apart from this successful intervention, the example contains a semantic error of substitution. Although the substitution ("if they work hard") is plausible based on the context preceding this segment, it departs from the original idea unit, which was "pero necesitas superar algunos obstáculos mentales del síndrome del impostor" ("but you need to overcome some of the mental hurdles of the imposter syndrome"). The output provides evidence of the complexity of live processing in respeaking: it shows an example of successful EE (structural re-expression) mixed with a substitution error, the latter being subject to penalisation under the NTR model.

4.3. Compensation

Working with live input provides respeakers with opportunities to change the sequence of information presentation if such a change is possible, i.e., in situations when the chronology of presentation is not crucial. Compensation consists in providing missing information from a previous idea unit (which can be either its part or even an entire idea unit) in later target idea units. Although typically compensation spans adjacent idea units, it is also possible for language professionals to compensate for content provided a few idea units earlier.

Participant comments from the TAPs confirm the existence of the category and reveal the rationale behind the use of compensation, as in examples 8 and 9 below, pertaining to a case where, in a multiple-speaker setting, a question was compensated for in the answer:

Participant 173: "This is another example where in trying to finish the previous idea, I actually missed the question, so my solution for that has been to ignore the question, and I just tried to speak the answer in a coherent way."

Table 2

Example 8 (intralingual respeaking)

Source	Target
And, of course, we also have the the pension pay gap which is very very serious. The current gender pension gap in Europe stands as over double the gender pay gap at 35.7%	The current gender pension gap in Europe **is serious** and stands at over double the gender pay gap, at 35.7%.

Table 3

Example 9 (interlingual respeaking Spanish > English)

Source	Target
Entonces, ¿significa esto que las personas que se dedican al rehablado intralingüístico pues que no tienen experiencia como intérpretes pueden dedicarse al rehablado interlingüístico? BT: *So, does this mean that people who perform intralingual respeaking, since they do not have experience as interpreters, can still perform interlingual respeaking?*	**I would say that people that haven't worked a simultaneous interpreters can work as interlingual respeakers, but there are obstacles to overcome.**
Yo diría que sí, absolutamente pero necesitas superar algunos obstáculos mentales del síndrome del impostor BT: *I would say yes, absolutely, but you need to overcome some of the imposter syndrome's mental hurdles*	

Examples 8 and 9 show that source information can be compensated if the respeaker's working memory allows it. Example 9 also shows that EEs can co-occur within the same segment – in the case of this example, compensation with structural re-expression (the latter already seen in Example 7). Importantly, due to its non-chronological nature, compensation is also uniquely human.

Table 4

Example 10 (intralingual respeaking)

Source:	Target:
Moi, je sais que vous souhaitez en savoir plus sur différent pays, *BT:I know that you would like to know more about the different countries*	
mais le cas de la France est particulièrement intéressant. *BT: but the case of France is particularly interesting*	**In France, it is particularly interesting and differs from other countries.**

Source: Authors' own elaboration.

Finally, in the target segment from Example 10, we notice how the content from the second source cell (representing a different idea unit) is merged with the previous idea in a plausible way that does not distort meaning. One may argue that this was possible because the respeaker started from the idea they heard last and then used their memory to integrate the preceding content.

5. Concluding Thoughts: Applicability of the Taxonomy and Further Challenges

The examples presented in this paper demonstrate that live spoken content in respeaking can undergo various positive changes implemented in real-time through human agency. While strategic reformulation has been thoroughly investigated in Interpreting Studies, there has been less systematic emphasis in the context of respeaking. To contribute to a more fine-grained definition of EEs, we have grouped these into three main categories: condensation, re-expression, and compensation. The extent of these changes depends on factors like the respeaker's briefing, original speech features (speed, complexity, genre), mastery of the technique, and individual preferences influenced by one's unique disposition, experience, and training. Importantly, the categories we have proposed are not intended as "watertight silos": language, and spoken language in particular, can cut across the boundaries we have delineated. Consequently, sometimes there will be borderline cases (e.g., between semantic and lexical level re-expressions), since these strategic behaviours can co-occur, as shown in the examples. Nevertheless, the main objective of the taxonomy is to provide guidance as to the types of positive editing that humans can implement in live settings. Such guidance seems necessary, as the impact of EEs may extend over several interrelated areas.

Firstly, it is critical for the accuracy evaluation of live subtitles to be able to distinguish between error categories on the one hand, and correct/effective editions on the other. While the former has been

clearly categorised into error types in the NER and NTR models, it is necessary to understand the range of different phenomena that fall into the latter (broad) category. This understanding is crucial, particularly since, as our study demonstrated, EEs are significant predictors of accuracy. Through our taxonomy, we hope to have shed more light on this aspect of evaluation, establishing functional links with existing categories within the NTR model. Furthermore, as subtitle accuracy evaluation is still a manual process, the humans trained to perform it can use the proposed classification of EEs to introduce more rigour and consistency to their efforts (this is particularly important when large pools of data are evaluated by teams of experts).

Another impact area refers to future (semi-)automation of subtitle accuracy evaluation: although accuracy evaluation with NER/NTR models is predominantly manual, labour-intensive and time-consuming, there are already attempts at (semi-) automating the process (Alonso-Bacigalupe & Romero-Fresco, 2023b) thanks to the analytical capacity of Large Language Models (LLMs). The authors do not reveal the exact prompting sequence that led to the creation of their prototype for semi-automated NER analysis (the "NER Buddy" – a system that uses the capacity of prompted LLM to detect and grade errors in subtitles). However, it is our assumption that a distinction between what constitutes a correct and incorrect edition (in the pro-active meaning of the term) must be somehow ingrained in the prompts. Given the preliminary results for the human-driven workflow (live subtitles delivered by a human respeaker working with speech recognition), it seems there is ample space for further elaboration of the chain-of-thought prompting implemented by the authors of the prototype. To elaborate on the prompts, we need (among other aspects) a more precise definition and exemplification of what constitutes effective editing. The impact of effective editing on the overall accuracy of live subtitles is not to be underestimated, as in some languages, text reduction in live subtitling through respeaking can be very significant and reach as much as 40% (e.g., Fresno & Romero-Fresco, 2022; Sandrelli, 2020).

Moreover, a deeper understanding of the effective editing phenomena in respeaking is crucial for training and upskilling. In the short and medium run, effective training for human professionals must include the application and reinforcement of live editing skills, facilitated by clear definitions and examples introduced at early stages of instruction. In the long run, the training of automated solutions for live captioning must also include some insights from our knowledge of effective editing, under the assumption that future automated systems will be required to produce live subtitles that are at least as readable and accessible as live subtitles from the human-driven workflow. Finally, a clearer picture of what constitutes effective editing has the potential to impact industry standards, including the consensus around what represents accessible and readable subtitles.

As indicated in the title of this paper, with the taxonomy proposed here, we hope to "move towards" a deeper understanding of EEs in live subtitling. Based on our analyses and insights gained during the development of the taxonomy, we are also able to point at further research directions and areas. Firstly, an experimental re-calculation of NER/NTR scores with positive values for EEs can be performed, preferably on large databases such as the SMART database. It may turn out that the positive aspect of EEs can partially offset the negative penalty scores to increase the overall

acceptability of the subtitles. Secondly, and related to the previous point, there appears to be a need for a reception study with materials featuring many EEs on the one hand and few EEs on the other. This could reveal end users' perceptions of EE-rich texts and lead to broader conclusions regarding optimal approaches to respeaking and/or automation. Thirdly, the proposed taxonomy can be experimentally turned into prompts for LLMs to find out if the detection and description of EEs, a time-consuming exercise, can be outsourced to an automated system. This paper hopes to inspire further investigation in all three areas.

Funder information

This study was part of the SMART project (*Shaping Multilingual Access Through Respeaking Technology*, ES/T002530/1, 2020–2023) funded by the Economic and Social Research Council UK and run at the University of Surrey's Centre for Translation Studies.

Acknowledgements

This paper contains references to analytical data co-created by former SMART Project team members. We would like to extend a special thanks to Anna-Stiina Wallinheimo for her input and guidance in statistical calculations, Zoe Moores for her contribution at the initial stages of taxonomy development as well as TAP analysis, and all evaluators who have been instrumental in analysing this vast dataset, namely Hayley Dawson, Eimee Brown, Katia Bulgarelli, Giulia Chialastri, Megan Stockwell, Francesco Saina, Sabrina Toscani.

References

Alonso-Bacigalupe, L., & Romero-Fresco, P. (2023a). Interlingual live subtitling: The crossroads between translation, interpreting, and accessibility. *Universal Access in the Information Society*, *1*(15). https://doi.org/10.1007/s10209-023-01032-8

Alonso-Bacigalupe, L., & Romero-Fresco, P. (2023b). The application of artificial intelligence-based tools to intralingual resleaking: the NER Buddy. In G. Corpas Pastor & C. M. Hidalgo-Ternero (Eds.), *Proceedings of the International Workshop on Interpreting Technologies SAY IT AGAIN* (pp. 9–15). Incoma.

Bang, J.-U., Choi, M.-Y., Kim, S.-H., & Kwon, O.-W. (2020). Automatic construction of a large-scale speech recognition database using multi-genre broadcast data with inaccurate subtitle timestamps. *IEICE Transactions on Information and Systems,* E103.D, 406–415. https://doi.org/10.1587/transinf.2019EDP7234

Bartłomiejczyk, M. (2006). Strategies of simultaneous interpreting and directionality. *Interpreting, 8*(1), 149–174.

Chmiel, A., Lijewska, A., Szarkowska, A., & Dutka, Ł. (2017). Paraphrasing in respeaking – comparing linguistic competence of interpreters, translators and bilinguals. *Perspectives: Studies in Translation Theory and Practice, 26*(5), 725–744.

Davitti, E., & Sandrelli, A. (2020). Embracing the complexity: A pilot study on inter-lingual respeaking. *Journal of Audiovisual Translation, 3*(2), 103–139.

Davitti, E., & Wallinheimo, A.-S. (2024*). Shaping multilingual access through respeaking technology* (Project Data, 2021) [Dataset]. UK Data Service. 10.5255/UKDA-SN-856687

Eugeni, C. (2008). A sociolinguistic approach to real-time subtitling: Respeaking vs. shadowing and simultaneous interpreting. *English in International Deaf Communication,* 72, 357–382.

Field, A. (2024). *Discovering statistics using IBM SPSS statistics* (6th ed.). Sage.

Fresno, N., & Romero-Fresco, P. (2022). Strengthening respeakers' training in Spain: The research-practice connection. *The Interpreter and Translator Trainer, 16*(1), 96–114. https://doi.org/10.1080/1750399X.2021.1884442

Gumul, E. (2017). *Explicitation in simultaneous interpreting: A study into explicitating behavior of trainee interpreters*. Wydawnictwo Uniwersytetu Śląskiego.

Hunt, M. J. (1990). Figures of merit for assessing connected word recognisers. *Speech Communication,* 9, 239–336.

Jones, R. (1998). *Conference interpreting explained*. St. Jerome Publishing.

Kaplan, A., Skogstad, A. L., & Girshick, M. (1949). *The prediction of social technological events*. Rand Corp. P-93.

Kalina, S. (1994). Analyzing interpretation performance: Methods and problem. In C. Dollerup & A. Lindegaard (Eds.), *Teaching translation and interpreting 2: Insights, aims & visions* (pp. 219–225). John Benjamins.

Kohn, K., & Kalina, S. (1996). The strategic dimension of interpreting. *Meta, 41*(1), 118–138.

Korybski, T., Davitti, E., Orăsan, C., & Braun, S. (2022). A semi-automated live interlingual communication workflow featuring intralingual respeaking: Evaluation and benchmarking.

In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (pp. 4405–4413). European Language Resources Association (ELRA).

Luyckx, B., Delbeke, T., Van Waes, L., Leijten, M., & Remael, A. (2010). Live subtitling with speech recognition: Causes and consequences of text reduction. *Across Languages and Cultures, 14*, 15–46.

Moser-Mercer, B. (2000). Simultaneous interpreting: Cognitive potential and limitations. *Interpreting, 5*(2), 83–94. https://doi.org/10.1075/intp.5.2.03mos

Pöchhacker, F. (2003). *Introducing interpreting studies* (1st ed.). Routledge.

Pöchhacker, F. (2009). Issues in interpreting studies. In J. Munday (Ed.), *The Routledge companion to translation studies* (pp. 128–140). Routledge.

Pöchhacker, F., & Remael, A. (2019). New efforts?: A competence-oriented task analysis of interlingual live subtitling. *Linguistica Antverpiensia, New Series: Themes in Translation Studies,* 18, 130–143.

Riccardi, A. (1996). Language-specific strategies in simultaneous interpreting. In C. Dollerup & V. Appel (Eds.), *Teaching translation and interpreting 3: New horizons* (pp. 213–221). John Benjamins.

Riccardi, A. (2005). On the evolution of interpreting strategies in simultaneous interpreting. *Meta, 50*(2), 753–767.

Robert, I. S., & Remael, A. (2017). Assessing quality in live interlingual subtitling: A new challenge. *Linguistica Antverpiensia New Series – Themes in Translation Studies*, 16, 168–195.

Romero-Fresco, P. (2011). *Subtitling through speech recognition: Respeaking*. St. Jerome.

Romero-Fresco, P. (2016). Accessing communication: The quality of live subtitles in the UK. *Language & Communication*, 49, 56–69.

Romero-Fresco, P. (2020). The accessible filmmaker and the global film. *MonTI. Monographs in Translation and Interpreting,* 12, 381–417. https://doi.org/10.6035/MonTI.2020.12.13

Romero-Fresco, P., & Pérez, J. M. (2015). Accuracy rate in live subtitling: The NER model. In J. Díaz Cintas & R. Baños (Eds.), *Audiovisual translation in a global context: Mapping an ever-changing landscape* (pp. 28–50). Palgrave Macmillan.

Romero-Fresco, P., & Pöchhacker, F. (2017). Quality assessment in interlingual live subtitling: The NTR model. *Linguistica Antverpiensia, New Series: Themes in Translation Studies,* 16, 149–167.

Sandrelli, A. (2020). Interlingual respeaking and simultaneous interpreting in a conference setting: A comparison. In *Technology in Interpreter Education and Practice* [Special issue]. *Intralinea*.

Tang, F., & Li, D. (2016). Explicitation patterns in English-Chinese consecutive interpreting: Differences between professional and trainee interpreters. *Perspectives: Studies in Translatology, 24*(2), 235–255. https://doi.org/10.1080/0907676X.2015.1040033

Wallinheimo, A.-S., Evans, S., & Davitti, E. (2023). Training in new forms of human-AI interaction improves complex working memory and switching skills of language professionals. *Frontiers in Artificial Intelligence,* 6. https://doi.org/10.3389/frai.2023.1253940

Wang, B. (2012). A descriptive study of norms in interpreting: Based on the Chinese-English consecutive interpreting corpus of Chinese Premier press conferences. *Meta, 57*(1), 198–212. https://doi.org/10.7202/1012749ar

Zue, V. W., Chen, S. F., & Glass, J. R. (1990). Error rates in automatic speech recognition and understanding. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(12), 1953–1960. https://doi.org/10.1109/29.60107