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

# Assessing quality in live interlingual subtitling: A new challenge

**Isabelle S. Robert**

Universiteit Antwerpen, Belgium
isabelle.robert@uantwerpen.be

**Aline Remael**

Universiteit Antwerpen, Belgium
aline.remael@uantwerpen.be

*Quality-assessment models for live interlingual subtitling are virtually non-existent. In this study we investigate whether and to what extent existing models from related translation modes, more specifically the Named Entity Recognition (NER) model for intralingual live subtitling, provide a good starting point. Having conducted a survey of the major quality parameters in different forms of subtitling, we proceed to adapt this model. The model measures live intralingual quality on the basis of different types of recognition error by the speech-recognition software, and edition errors by the respeaker, with reference to their impact on the viewer's comprehension. To test the adapted model we conducted a context-based study comprising the observation of the live interlingual subtitling process of four episodes of* Dansdate*, broadcast by the Flemish commercial broadcaster VTM in 2015. The process observed involved four "subtitlers": the respeaker/interpreter, a corrector, a speech-to-text interpreter and a broadcaster, all of whom performed different functions. The data collected allow errors in the final product and in the intermediate stages to be identified: they include when and by whom they were made. The results show that the NER model can be applied to live interlingual subtitling if it is adapted to deal with errors specific to translation proper.*

## 1. Introduction

Quality-assessment models for live interlingual subtitling with speech recognition are virtually non-existent and designing such a model is a complex undertaking. One reason for this is the relative novelty of the translation mode, which means that research is scarce and both practical experience and data are limited. Another reason is the hybrid character of live subtitling. While being a translation mode in its own right, it shares features with its closest variant, intralingual live subtitling (Romero-Fresco, 2011), but also with more traditional, pre-prepared forms of interlingual and intralingual subtitling (Díaz Cintas & Remael, 2007), and even with simultaneous interpreting (SI) (Pöchhacker, 2016). All three of the abovementioned subtitling modes produce subtitles. However, their production processes differ greatly, and each has its own impact on the type of end-product or subtitle it yields. In addition, the target-user groups of all the subtitling modes vary, resulting in different expectations and demands. In a nutshell, one could summarize the different production processes as follows: in the pre-prepared intralingual and interlingual mode the subtitles are produced with dedicated subtitling software through a non-live rephrasing or translation process in post-production; in live intralingual subtitling augmented by speech recognition the subtitles are now produced mainly with speech-to-text software through a live form of rephrasing, and in live interlingual subtitling this live feature is combined with a variant of simultaneous interpreting. As far as the process is concerned, live in*ter*lingual subtitling therefore shares the most common ground with live in*tra*lingual subtitling. Both modes require a form of "respeaking", a procedure that has been described as follows:

> In an average live subtitling session, one person watches and listens to the television program[me] as it is broadcast live. Wearing a headset, he or she simultaneously repeats, paraphrases or "respeaks" what is being said. "Respeaking" is the currently used term for this

> activity but … it is a bit of a misnomer since the subtitler does more than respeaking alone. Nevertheless, this so-called respeaker speaks directly to a speech recognizer, which produces a draft subtitle. [In the case of intralingual respeaking in Flanders] errors that are made by the respeaker or by the speech recognizer itself are corrected before they are put on the air. This can be done either by the respeaker (the Mono-LS-model) or by an additional editor who will quickly correct the output of the speech recognition program before the subtitles are broadcast (the Duo-LS-model). (Remael, Van Waes & Leijten, 2014, p. 124)

One important challenge concomitant with the live subtitling process is the so-called "delay" that occurs between the moment the respeaker receives the audio input through their headphones and the moment the final subtitle appears on the television screen (see, for example, Van Waes, Leijten & Remael, 2013, for more details). However, given that all the modes yield subtitles, live interlingual also shares the features and therefore the quality requirements of both of the other modes. A central question is the extent to which the challenges of the different live production processes have an impact on the quality that can be expected and deemed acceptable by broadcasters and viewers. However, to answer this question, one must first know whether and how this quality can be assessed and which parameters must be taken into account. The aim of this article is therefore fourfold:

1. to review briefly the main quality parameters or criteria used in subtitling practices with which live interlingual subtitling shares some common ground, that is, intralingual and interlingual pre-prepared subtitling, and intralingual live subtitling with speech recognition, with a brief excursion into SI;
2. to develop a tentative quality-assessment model for interlingual live subtitling based on this review;
3. to test the usability of the model in a case study;
4. to assess the results produced by means of the procedure currently used at VTM, the main commercial Flemish broadcaster, and formulate suggestions for further research.

In section 2 we briefly survey central quality issues and requirements in pre-prepared subtitling and intralingual live, with a particular focus on the concepts developed in the now widely used Named Entity Recognition (NER) model (Romero-Fresco, 2016). The model was designed for quality assessment in intralingual live subtitling. The subsection ends with a brief excursion into quality issues in SI. Section 3 presents a tentative quality-assessment model for live interlingual subtitling, based mainly on the NER model. The model aims to categorize and quantify the errors that occur in the different stages of the process, with the NER model as its starting point. Section 4 is dedicated to methodology, that is, an observational case study. This consists of the first descriptive study of the live interlingual English–Dutch subtitling process as it was carried out for *Dansdate*, a live talk show broadcast by the commercial Broadcaster VTM (Flanders, Belgium) in 2015. This section also includes a report on data collection and data analysis. Section 5 presents and discusses the results and offers a first evaluation both of the model and of the current process and product, within the limitations of the case study. The concluding comments in section 6 contain some suggestions for best practice and further research.

## 2. Issues of quality in subtitling and simultaneous interpreting: main challenges

### 2.1. Interlingual and intralingual pre-prepared subtitling

The central issues determining the quality of pre-prepared interlingual subtitling have been detailed in Díaz Cintas and Remael (2007), Kuo (2014) and, more recently, in Robert and Remael (2016). They can be subdivided into form-related and content-related quality issues, even though the two are strongly connected. The most important form-related issues include:

- formatting (segmentation and layout);
- spotting (synchronicity with the spoken text and reading speed), and
- readability (in terms of font type and size).

The most important content-related issues or translation-related issues are shown in Table 1 (adapted from Robert and Remael (2016) and Carrol and Ivarsson (1998)).

Table 1: Content-related quality parameters in subtitles

| | |
|---|---|
| Content and translation (including accuracy, completeness, logic) | There must be a close correlation between film dialogue and subtitle content; source language (SL) and target language (TL) should be synchronized as far as possible. |
| Grammar, spelling and punctuation | The language should be grammatically correct since subtitles serve as a model for literacy.<br>Simple syntactic units should be used. |
| Readability (i.e., ease of comprehension and coherence between individual subtitles) | When it is necessary to condense dialogue, the text must be coherent. |
| Appropriateness (socio-cultural features of the audience) | Translation quality must be high with due consideration of all idiomatic and cultural nuances.<br>The language register must be appropriate and correspond to locution. |

Broadly speaking, the above parameters are also applicable to intralingual pre-prepared subtitling for the Deaf and Hard of Hearing (SDH) but additional factors must be taken into account in view of the different (and also diverse) needs of the SDH target audience. The degree to which and the manner in which the additional parameters are applied in practice differ across countries; however, the challenges themselves are well known. A fairly recent study on SDH in Spain (Pereira 2010), for instance, identified:

- speaker identification (e.g., through the use of colours but also through subtitle placement on screen,
- high synchronicity with speech and information on screen, and the written rendering of oral speech);
- time of display on the screen or reading speed (which may have to be slower in this Audiovisual Translation (AVT) mode);
- signs rendering aural information (orthotypographical criteria such as suspension points to signal hesitation in speech and didascalies (i.e., "stage directions") to represent aural information, such as sighs, music and noises).

### 2.2. Intralingual live subtitling

In Flanders, intralingual live subtitling was introduced to meet the quota imposed on the public broadcaster (100 per cent subtitling of Dutch programmes) and to ensure the subtitling of live news on commercial channels. This means that the main target group of this AVT mode is the SDH audience. In other words, achieving a combination of the requirements for producing quality subtitles in the interlingual and intralingual pre-prepared modes should, ideally, be the aim. Practice, however, has shown that this is virtually impossible, despite attempts having been made to respect reading speed through editing and colours added to identify speakers (see, for example, Van Waes et al., 2013; Romero-Fresco, 2011). The major challenge to the production process lies in the live multitasking

required of the respeaker, who must perform manual actions on a keyboard while rephrasing (in edited or verbatim condition,[1] as the case may be). This tends to result in the production of errors and the delay referred to in section 1.

Reconsidering the formal features listed as important in pre-prepared subtitling, the following features are challenged most in a live context because of the multitasking involved:

- formatting (segmentation and layout);
- spotting (synchronicity between the spoken text and reading speed).[2]

Although the importance of careful formatting (i.e., segmentation based on semantic criteria) has been demonstrated (see, e.g., Díaz Cintas & Remael, 2007; Rajendran, Duchowski, Orero, Martínez & Romero-Fresco, 2013), respeakers tend to follow the speech rhythm of the speaker, trying to combine the speaker's natural segmentation with their intuitive feel for subtitle length and reading speed. This results in greater variation in subtitling length and layout (for examples see Van Waes et al., 2013). As for spotting and synchronicity, the delay that is typical of the live subtitling process makes full synchronicity impossible, unless the respeakers are given a head start on broadcasting the programme in the form of what is known as a "broadcast delay" (of a variable number of minutes).

However, it is the translation-related or content-related features that concern us most in the current article, and especially the impact of live rephrasing or translating and condensation on:

- content and translation;
- grammar, spelling and punctuation;
- readability, that is, coherence, and
- appropriateness (cf. Table 1).

Previous research by Van Waes et al. (2013) has demonstrated that the speed of delivery (number of words per minute) has an effect on condensation and that condensation in live intralingual subtitling can vary greatly. Nevertheless, this does not necessarily mean a great loss of content. Nor, in other words, does it mean, as in the case of pre-prepared subtitles, that shorter (live) subtitles automatically mean that translation errors have been made or that important information has been omitted. That is why assessing the quality of subtitle content requires human intervention, whether the subtitles to be assessed are intralingual or interlingual, live or pre-prepared.

To return to intralingual live subtitling, a number of studies have attempted categorizations of the content-related errors that occur in such subtitles. They are also surveyed in Romero-Fresco (2016), who discusses both the latest version of his NER model and its application by Ofcom in the United Kingdom, plus its increasing popularity among broadcasters worldwide. Since this is the model we aim to test for interlingual live, NER is the subject of the next subsection

### 2.3. The NER model

An important feature of the NER model is that it takes account of typical subtitling characteristics such as correct editions (CE; see below), which standard Word Error Rate (WER) calculations, traditionally used to assess speech-to-text engines, do not take into account. In other words,

> [t]he model draws on the basic principles of WER calculations and factors in different severity of errors while accounting for different degrees of reduction in the subtitles and highlighting the need for human intervention in the assessment of subtitle quality. (Romero-Fresco, 2016, p. 57)

More concretely, NER classifies content-related errors according to their severity, attributes scores to them and feeds them into the formula below:

$$\text{Accuracy rate} = [(N - R - E)/N] \times 100$$

In this formula, N stands for the number of respoken words in the subtitles and E stands for edition errors (EEs). Edition errors are usually caused by the strategies applied by the subtitler, for instance the omission of information, but also the introduction of wrong information. EEs are identified by comparing the subtitles against the transcript of the audio. R stands for recognition errors (REs), which are misrecognitions caused by mispronunciations or mishearing, or by the specific technology used to produce the subtitles. EEs and REss are classified as serious, standard or minor, scoring 1, 0.5 and 0.25, respectively:

- A serious error (penalized by 1) changes the meaning of the original audio, introducing factual mistakes or misleading information that can nevertheless make sense in the context and therefore go unnoticed.
- Standard errors (penalized by 0.5) involve omitting an information unit from the source text – often a fully independent idea unit or a dependent idea unit that renders the remaining unit nonsensical – which is noticed by the viewer.
- Minor errors (penalized by 0.25) nevertheless allow viewers to understand the broadcast; they involve omissions that do not affect the meaning of the remaining unit (they simply produce less information or fewer spelling errors, for instance).

To conclude, the NER model also distinguishes correct editions (CEs), which may involve condensation or rephrasing but which do not affect the overall meaning of the source text rendering and are therefore not included in the formula.

Romero-Fresco's assessment of the model has also determined the threshold that must be reached for subtitles to meet acceptable to very good accuracy rates, translated into the following percentages (see Table 2).

Table 2: Accuracy rate and corresponding quality level

| Accuracy rate (%) | Quality |
|---|---|
| < 98 | Substandard |
| 98–98.49 | Acceptable |
| 98.5–98.99 | Good |
| 99–99.49 | Very good |
| 99.5–100 | Excellent |

The questions arising from the application of the NER model are these: How can this model be applied to interlingual live subtitling, combining its classification of errors for intralingual live with the translation-related classification of quality parameters distinguished above for pre-prepared interlingual subtitling? And must interpreting-related parameters also be taken into account?

### 2.4. Main quality parameters in simultaneous interpreting

An in-depth review of the quality-assessment models or of the quality criteria applied in SI is beyond the scope of this article. However, two fairly recent studies specifically dedicated to quality criteria in interpreting revealed the use of quality criteria similar to those listed for subtitling, insofar as they can be applied to an oral as opposed to a written form of translation.

In his study on the expectations of interpreting services users, Kurz (2001) investigated the significance they attribute to eight quality criteria for interpretation: accent, voice, fluency, logical cohesion, sense consistency, completeness, grammar and terminology. Similarly, in a study on

interpretation services quality as perceived by the providers of such services, Zwischenberger (2010) listed 11 criteria distributed across three categories: content-related criteria (sense consistency, logical cohesion, completeness), form-related criteria (correct terminology, correct grammar, appropriate style) and delivery-related criteria (fluency of delivery, lively intonation, pleasant voice, synchronicity and native accent). In other words, in both studies, many content-related and form-related criteria are similar to those applied in subtitling (see sections 2.1 and 2.2). As far as delivery-related criteria are concerned, only synchronicity seems to be a shared feature. Voicing is important in both interpreting and live subtitling but in different ways. For the speech-to-text software to function efficiently, respeakers adopt a clearly articulated but monotone speech pattern, since the computer is their first target "audience". The interpreters' audience, by way of contrast, consists of human beings, which means the interpreters' output must be lively and must be delivered with the appropriate intonation. Nevertheless, for content- and form-related criteria, as described for pre-prepared subtitling, the same central items recur. We will return to these in our case study below.

## 3. A quality-assessment model for live interlingual subtitling

Drawing on the main interlingual translation quality parameters presented in section 2, and taking the NER model for intralingual live subtitling as its main source of inspiration, the next aim of this study is to design a first model for the quality assessment of live interlingual subtitling.

The typology the study proposes encompasses the same types of error as the NER model, that is, REs, EEs and CEs. Contrary to the NER model, however, it distinguishes between two types of impact that an error can have on the viewers: that on the comprehension of the viewers and that on the perception of the viewers, that is, their perception of the quality the producer, distributor and/or the broadcaster deliver in the translation.

Contrary to what is customary when using the NER model, all REs have been rated as standard (−0.5) since substantial analysis of our data from *Dansdate* revealed that the majority of the REs had – or would have had, had they not been corrected – a "standard effect" on the viewer's comprehension, as defined in the NER model. That is, they were errors that are identified as errors but for which it is not always easy to understand what was originally meant, errors which do not create a new meaning but result in the omission of an information unit from the original text, or errors which disrupt the flow or meaning of the original and often cause surprise (Romero-Fresco, 2016, p. 59).

As far as EEs are concerned, the typology has been adapted for interlingual subtitling in the sense that it now also includes typical translation errors that are not necessarily associated with intralingual paraphrasing, based on a thorough analysis of all the translation shifts and errors occurring in our corpus. In addition, we tried to take the impact of visually supplied information into account when assessing the appropriateness of translations. The result is a typology that includes errors and shifts that are generally not found in intralingual subtitling. It is summarized in Table 3. As in the NER model, our typology distinguishes between minor (0.25), standard (0.5) and serious errors (1). However, we introduced an additional, more fine-grained level, that is, standard+ errors (0.75) (see below).

Besides, our weighting depends on the type of effect the error is thought to have on the reader, that is, one related to perception, on the one hand, and another related to comprehension, on the other.[3]

Table 3: Error typology for interlingual live subtitling

| Type | Impact | Weight | Label | Description |
|---|---|---|---|---|
| RE | Comprehension | Standard | RE | Rcognition error |
| EE | Perception | Minor | Punctuation | Missing or inappropriate punctuation |
| | | Minor | Spelling | Spelling mistake |
| | | Minor | Idiom, register | Non-idiomatic formulation in Dutch (generally an interference) or inappropriate register |
| | | Standard | Grammar | Grammar mistake |
| EE | Comprehension | Minor | Slight shift | Slight shift in meaning |
| | | Minor | Small omission | Omission with little meaning loss. Examples: cohesion between subtitles, missing direct object modifier, missing second direct object, missing subject attribute modifier, missing indirect object, missing adverbials (time, place, emphasis, degree) |
| | | Standard | Compensated wrong meaning | Wrong meaning compensated by the visual content |
| | | Standard+ | Omission | Missing clause |
| | | Standard+ | Nonsense | Formulation making no sense |
| | | Serious | Wrong meaning | Wrong meaning which is not compensated by the visual content |
| CE | No impact | Zero | Omission without meaning loss | Omission of superfluous, redundant or transparent information. Examples: repetitions, hesitations, discourse markers, adverbials (modality), coordinating conjunctions |

As in the NER model, punctuation and spelling errors are considered minor, since they have no effect on comprehension, but they might convey a negative impression and therefore might have a slightly negative impact on the perception the users have of the quality provided by the producer and/or the broadcaster. Grammar mistakes, however, are thought to have a higher negative impact than punctuation errors (at least, higher than all the punctuation errors encountered in our corpus) and have therefore been rated as standard.

Moreover, the aforementioned type of effect on the user, that is, the negative impact of an error on the perception of quality by the user, plays an important role in the weighting of errors. Therefore, a grammatical error (normally EE standard) that can be expected to have an impact on comprehension is considered “an error with comprehension impact” and can be categorized as

standard+ (e.g., formulation making no sense). The same holds for a spelling error: if a spelling error (which will often be only an EE minor) leads to a new, erroneous meaning, it can be considered as a slight shift or even a wrong meaning (EE serious).

The distinction between omission with meaning loss (EE) and without meaning loss (CE) in the NER model has also been maintained. As far as omissions with meaning loss are concerned, the typology distinguishes between minor omissions and standard+ omissions. A standard+ omission corresponds to the omission of a whole clause. The data analysis showed one particular type of omission to be less serious than a serious error but more serious than a standard error. It is less serious than a serious error because it does not create a new meaning that would go unnoticed. It is more serious than a standard error because it does generally remain under the radar without supplying fundamentally wrong information, but it nevertheless does deprive the user of a major piece of information.

The typology also includes errors related to translation "proper", that is, nonsense, wrong meaning and slight shift. Nonsense is considered to be standard+, since it will be noticed by the viewer; it is therefore less serious than a wrong meaning that goes unnoticed, which is considered as a serious error (weight = 1). Finally, slight shifts count as minor errors.

Section 4 (methodology) details how the new typology was applied to the data analysis. The analysis was carried out by both authors, ensuring maximal interrater reliability (but see our conclusions).

## 4. Methodology: case study

The methodology selected to test the usability of the adapted NER model for interlingual live subtitling is a case study, defined as "an empirical inquiry that investigates a contemporary phenomenon in depth and within its real-life context, especially when the boundaries between the phenomenon and the context are not clearly evident" (Yin 2009, cited by Saldanha & O'Brien, 2013, p. 207).

The case study research type was chosen for several reasons. First, although live subtitling with speech recognition has a long tradition in Flanders, it is mostly intralingual, with live interlingual subtitling being restricted to selected programmes that feature a foreign-language speaker. In other words, it represents only a small proportion of the workload of professional subtitlers in Flanders, and it is largely restricted to the public broadcaster, VRT, and one major commercial channel, VTM. As a result, best practices have not yet been determined, which means that at this stage the observation and assessment of current practice in a real-life context is preferable to the experimental alternative. Indeed, both VRT and VTM are themselves still experimenting with different set-ups for interlingual live subtitling and are interested in assessments involving their current methods. Second, although case-study research focuses on one particular case, which means generalizations are impossible, it can make contributions to knowledge beyond the particular in three different scenarios: (1) for exploring questions of how and why, (2) for hypothesis generating (as opposed to hypothesis testing) and (3) for testing the viability of a theoretical framework. (Saldanha & O'Brien, 2013, p. 209)

These are precisely the objectives of the present study: (1) to determine how the live interlingual subtitling process is organized and, if necessary, why it should or how it could be reorganized; (2) to generate hypotheses about best practices in live interlingual subtitling, and (3) to test the viability of the NER model in the context of live interlingual subtitling (although the NER model is, strictly speaking, not a theoretical framework).

### 3.1 Live interlingual subtitling process at VTM: description

The live interlingual subtitling process at VTM is organized in the same way as the live intralingual subtitling process, with a few adjustments. The first adjustment is related to broadcast delay: although there are generally no delays in live intralingual subtitling, VTM often allows short broadcast delays for programmes requiring live interlingual subtitling. Such delays may range from a few seconds to

two minutes. For *Dansdate*, a broadcast delay of one minute was allowed. The second adjustment relates to the number and profile of the "subtitlers" responsible for the whole subtitling process, as detailed in the next paragraph. In addition, we need to point out that the subtitles of *Dansdate* were not meant for the SDH in the first place, but rather for a general Flemish audience, which means that speaker identification was not required. Moreover, there was only one foreign-language speaker. The other speakers were therefore not subtitled.

In the case of *live intralingual subtitling* the VTM team consists of one or two "subtitlers": a respeaker and, sometimes, a corrector,[4] depending on the perceived degree of difficulty of the programme. When the respeaker works alone, for programmes where spoken output is limited in quantity and is slow (e.g., a tennis match), the respeaker "respeaks" the oral input, the respoken output is converted into text by the voice-recognition software Dragon Naturally Speaking, and that text is automatically generated in the subtitling software, WinCaps. The respeaker then reads the subtitle, makes corrections where needed and broadcasts it. For more challenging programmes, the respeaker is assisted by a corrector, who makes additional changes in the subtitles where needed and is responsible for the broadcasting. In Flanders, block subtitling (as opposed to scrolling) is the norm, and the respeaking process is never verbatim; this means that non-relevant information, such as repetitions, hesitations, discourse markers, some adjectives, etc., are omitted, as they would be in pre-prepared subtitling.

In *live interlingual subtitling*, two additional "subtitlers" join the team: a speech-to-text interpreter (STI) and a broadcaster. Moreover, the respeaker is a trained interpreter, which is not necessarily the case in live intralingual subtitling. At VTM, the current interlingual live subtitling process can be described as follows (see Figure 1):

The first phase (phase 1A in Figure 1) involves the respeaker/interpreter (R/I) simultaneously interpreting the oral input from language A (English for *Dansdate*) into language B (Dutch). As we pointed out above, live interlingual subtitling shares some characteristics with SI.

The next phase (phase 1B) entails converting the interpreted output into text by Dragon Naturally Speaking and integrating that text into the WinCaps subtitling software. At this stage, just as in live intralingual subtitling, the R/I can read the subtitle and edit it (phase 1C).

However, since interlingual subtitling is more demanding than intralingual subtitling, the R/I makes far fewer corrections, the corrector having taken on this function (Phase 2). The corrector is assisted in their task by the STI. The corrector focuses mainly on the transfer from spoken Dutch to written Dutch and therefore mostly on REs, although they also correct EEs that are generally not related to the interlingual transfer, that is, to the translation process as such. The STI, in contrast, listens mainly to the original oral output in the foreign language (English) and compares it to the subtitle in the TL (Dutch). As in consecutive interpreting, the STI takes notes while listening to the English source speech. Since both the corrector and the STI share one and the same computer, the latter indicates any necessary changes to the former. The changes requested by the STI are generally related to EEs linked to the interlingual transfer.

Finally, the broadcaster, who listens to the broadcast output and not to the original studio output (i.e., the input with a one-minute delay) broadcasts the subtitles (phase 3).

This rather complex production process, which, as was pointed out above, is itself in an experimental phase, actually results in perfectly synchronous subtitles most of the time, thanks to the collaborative effort of the team and the one-minute delay.[5]

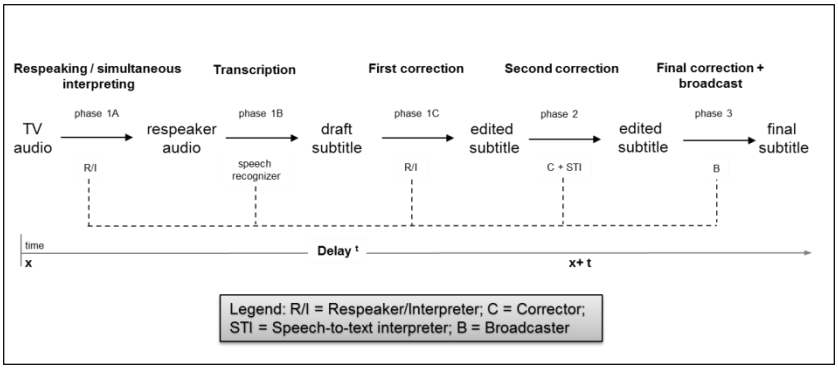


Figure 1: Live interlingual subtitling process at VTM

## 4.2. Data collection

### 4.2.1. Context of observation: *Dansdate* on VTM

*Dansdate* is an entertainment dance programme that was broadcast once a week on VTM from October to December 2014. It consisted of six episodes, but only four of these (episodes 3 to 6) were used for our study: the first episode was not broadcast live (it had a broadcast delay of half an hour) and when the second episode was aired, some of the technology put in place to record the whole subtitling process was not functioning properly.

In *Dansdate* a Flemish personality such as a famous actor, singer, politician or journalist performed a particular type of dance (e.g., tango, disco, walz, cabaret) with their real-life partner. The show started with six couples and from week three one couple was eliminated every week, with only two couples remaining in the final episode. The jury consisted of the well-known Flemish television actress and presenter, Francesca Vanthielen, the Belgian choreographer, Min Hee Bervoets, and Daniel Quinn Karaty, the American TV personality, actor, producer, dancer and choreographer. Karaty was also a judge on several versions of an earlier comparable production *So You Think You Can Dance*[6] which was a source of inspiration for *Dansdate*.

### 4.2.2. Observers

Our observers were two students preparing their master's theses about live interlingual subtitling under the supervision of the present authors. Their work was used as the starting point for this study. The two students decided to research this particular type of subtitling process because they were themselves freelance live subtitlers at VTM. In other words, they were familiar with all the team members, both the intralingual and the interlingual subtitling process at VTM, and the working conditions there. Consequently, their observations were neither fully detached nor participatory (for more information, see Saldanha & O'Brien, 2013, pp. 222–223). In both types of observation, a Hawthorne effect is generally inevitable: people act differently because they know they are being observed. However, in this case, both observers reported that they did not find the observed processes noticeably different from the subtitling processes they had previously participated in.

### 4.2.3. Participants

In total, seven people were observed, since it was not possible to have the same team for each episode. Participants C, D, E were part of the team three times, participants A, B and G twice, and participant

F just once. The functions they performed were distributed over the four observed episodes described in Table 4. Except for D, who always took on the role of broadcaster, the participants played different roles, although E was the R/I twice and C took on the role of the corrector twice.

Table 4: Distribution of the different functions per episode

| Episode | R/I | Corrector | STI | Broadcaster |
|---|---|---|---|---|
| DD3 | B | C | A | D |
| DD4 | E | C | F | D |
| DD5 | G | A | E | C |
| DD6 | E | B | G | D |

All the members of the subtitling department at VTM have a master's degree in Translation and/or Interpreting (or CI). The subtitling department is part of the news department and its main task is subtitling the news (three times/day) and producing both open and closed subtitles, mainly intralingually, but occasionally also interlingually, as explained in section 1. We did obtain some more information about the subtitlers' professional backgrounds and experience; however, this was too general to be of use in interpreting the data yielded by their respective performances. We return to this issue in the conclusions.

### 4.2.4. Types of data collected

As indicated above, the data collection was carried out for episodes 3 to 6. Different types of data were collected and these are summarized in Table 5. Inputlog is logging software developed at the University of Antwerp (http://www.inputlog.net/).

Table 5: Type and description of collected data

| | Type | Description |
|---|---|---|
| 1 | Video files | – interlingual subtitles and audio output in the SL (English) |
| 2 | Wincaps files | – subtitles in Dutch |
| 3 | Audio files (digital audio recorder) | – output of the respeaker in the TL (Dutch)<br>– conversation between the corrector and the STI<br>– retrospective team feedback[7] |
| 4 | Inputlog files | – from respeaker computer<br>– from corrector computer<br>– from broadcaster computer |

## 4.3. Data analysis

The data analysis was based on a reconstruction of the entire live interlingual subtitling process with the different types of data described in Table 5. An Excel working table with a column for each stage of the process was used to organize the data: (1) English oral input, (2) respoken output, (3) Dragon text in Wincaps, (4) text with changes by the corrector and/or STI, (5) text with changes by the broadcaster and (6) final broadcast text. The changes made by the respeaker, corrector/STI or

broadcaster could be identified, thanks to the General Analysis file of Inputlog. A simplified reproduction of our Excel working table is given in Table 6.

Table 6: Simplified reproduction of the Excel working table

| DD nb. | Sbt. nb. | English oral input | R/I output | Dragon conversion in WinCaps | Corrector | Broadcaster (final) | Back translation of the final subtitle |
|---|---|---|---|---|---|---|---|
| 4 | 38 | You guys didn't just get the steps and the style, you got the essence of what it is, | jullie hadden niet enkel de pasjes goed[,] maar ook de essentie[.] | [?] niet enkel [?] goed, maar ook de essentie. | je had niet enkel de pasjes<br>_____<br>onder de knie, maar ook de essentie. | je had niet enkel de pasjes<br>_____<br>onder de knie, maar ook de essentie. | You didn't just get the steps,<br><br>but also the essence. |

Legend: DD nb. = number of *Dansdate* episode; Sbt. nb. = number of subtitle. In column 4, elements between square brackets are punctuation marks dictated by the R/I. In column 5, elements between square brackets are REs by Dragon. In columns 6 and 7, the horizontal line means that the subtitle consisted of two lines.

In our draft Excel table, for each subtitle we indicated the number of words in the English oral input, in the respoken output and in the final broadcast text. We also indicated the type of error, who made the error and, where appropriate, who corrected the error, in separate columns. Errors were coded according to their type: taking NER as a basis, we distinguished between REs, EEs and CEs.

The coding of all the errors enabled us to calculate an accuracy rate per subtitle and per stage, following the formula from the NER model (see section 3). In other words, we were able to calculate four different accuracy rates for the same subtitle: one for **Dragon** (i.e., the accuracy rate of the subtitle if Dragon output had been broadcast without any correction), one for the **R/I** (i.e., the accuracy rate of Dragon, taking into account the corrections made by the respeaker), one for the **corrector/STI** taken together (i.e., the accuracy rate of the respeaker, taking into account the corrections made by the corrector and the STI, as well as the errors introduced at this stage), and one for the **broadcaster** (i.e., both the accuracy rate of the corrector, taking into account the corrections made by the broadcaster, and the errors introduced at this stage).

Table 7 is an illustration of this procedure. It is an excerpt from the score calculation for episode 3. Row labels correspond to the number of the subtitle. Table 7 indicates that there were no errors in subtitle 1. In row 4 (subtitle 4), nobody corrected the error (minor in this case). In row 5, corrections were made by the respeaker (taking the rating down from 3.25 to 2.75) and by the corrector (taking it down from 2.75 to 2). The broadcaster did not introduce any further corrections. Row 6 contains an example of an erroneous correction by the corrector, since his score is higher than the respeaker's score. As can be seen from the last row (total), the error score reduces at each step of the process: from 55 for Dragon to 28.5 for the broadcaster. The biggest improvement seems to occur between the interventions of the respeaker and the corrector (from 50.25 to 32.25). To calculate the accuracy rate, we used the NER formula, that is, we included the number of words for each subtitle in order to generate the accuracy rate in percentages. An example is shown in Table 8, with the accuracy scores corresponding to Table 7. Since nobody made an error in row 1 of Table 7 (zero score), the percentage in row 1 of Table 8 is 100.

Table 7: Examples of scores in absolute numbers

| Row labels | Score Dragon | Score RI | Score C/STI | Score B |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 4 | 0.25 | 0.25 | 0.25 | 0.25 |
| 5 | 3.25 | 2.75 | 2 | 2 |
| 6 | 0.75 | 0.25 | 0.50 | 0.25 |
| [...] | | | | |
| Total | 55.00 | 50.25 | 32.25 | 28.50 |

Table 8: Examples of scores in percentages

| Row labels | Score Dragon | Score RI | Score C/STI | Score B |
|---|---|---|---|---|
| 1 | 100.00 | 100.00 | 100.00 | 100.00 |
| 4 | 95.00 | 95.00 | 95.00 | 95.00 |
| 5 | 67.50 | 72.50 | 80.00 | 80.00 |
| 6 | 94.64 | 98.21 | 96.43 | 98.21 |

## 5. Results

In section 2, we briefly reviewed the existing quality-assessment parameters or criteria used in practices with which live interlingual subtitling shares some common ground in terms of process and product, that is, intralingual and interlingual pre-prepared subtitling, intralingual subtitling with speech recognition and SI. The aim was to develop a tentative quality-assessment model for interlingual live subtitling based on this review. That model was presented in section 3. As shown in section 4, and in particular in section 4.3, we were able to apply the newly designed model to our case study.

In the next section, we concentrate on three types of analysis: (1) the accuracy rate for each step of the process, for each episode separately; (2) the types of error made and corrected at each step and their impact on quality, and (3) the relation between the accuracy rate and some characteristics of the episode (e.g., number of words in SL).

### 5.1. Accuracy rate in *Dansdate*

As indicated in section 4.3, the accuracy rate was calculated for each subtitle of each episode for every step of the process. Consequently, an average accuracy rate could be calculated for every step of each episode. The descriptive statistics are summarized in Table 9.

Table 9: Accuracy rate for each step in the process, per episode

| Accuracy rate | Dragon<br>M<br>SD | Respeaker<br>M<br>SD | Corrector<br>M<br>SD | Broadcaster<br>M<br>SD |
|---|---|---|---|---|
| DD3 | 89.86 | 90.53 | 94.54 | 95.17 |
| | 10.40 | 10.34 | 9.22 | 9.09 |
| DD4 | 93.44 | 94.51 | 97.95 | *98.26 |
| | 7.86 | 7.40 | 3.63 | 3.34 |
| DD5 | 94.96 | 95.31 | 97.51 | 97.72 |
| | 7.06 | 6.81 | 5.80 | 5.67 |
| DD6 | 94.94 | 95.87 | *98.45 | **98.66 |
| | 8.70 | 8.43 | 3.66 | 3.53 |

Note: One asterisk means that the accuracy rate reached an acceptable level; two asterisks means that it reached a good level. All the scores without asterisks are substandard.

Table 9 shows that the average score for episode 3 (DD3), for every phase, remained substandard, even in the final product (broadcast subtitles). The same applies to episode 5 (DD5). However, in the fourth episode, the broadcast subtitles reached an acceptable level, while in the sixth episode an acceptable level was reached in the third phase and a good level in the last phase, that is, in the broadcast subtitles. All in all, the results reveal small differences from one phase to the next. On average, the difference between the Dragon accuracy rate and the respeaker accuracy rate is just 0.76; the difference between the respeaker accuracy rate and the corrector accuracy rate is higher, on average, at 3.06. As for the last phase, here the difference in the accuracy rate is again very low, at 0.34. The question therefore is whether the differences in accuracy rate are significant.

To answer that question, we conducted a Friedman test of comparison of several non-independent groups. We applied a non-parametrical test because the data were not normally distributed (significant Kolmogorov-Smirnov test of normality) and we considered the series of scores to be non-independent, not because the participants responsible for these scores are the same (they are not), but because the scores are related to one another, since each step and therefore each score depends on the previous one. Since the Friedman test revealed significance, which means that an overall effect was present, we conducted a similar parametrical test (ANOVA for repeated measures) that can reveal contrasts between levels – in our case, contrasts between the different phases in the process. The results of both tests are summarized in Table 10. The results relating to contrasts between the different phases are included in Table 11.

Table 10: Friedman and ANOVA tests (comparison of several non-independent groups)

| Friedman | DD3 | DD4 | DD5 | DD6 |
|---|---|---|---|---|
| N | 77 | 110 | 104 | 90 |
| Chi-Square | 123.07 | 147.43 | 103.39 | 64.62 |
| Df | 3 | 3 | 3 | 3 |
| Monte Carlo Sig. | .000 | .000 | .000 | .000 |
| ANOVA | | | | |
| F (Greehouse-Geisser) | 37.01 | 36.82 | 33.61 | 13.57 |
| Df | 3 | 3 | 3 | 3 |
| Sig. | .000 | .000 | .000 | .000 |

Table 11: Contrasts from one step to the next

| | DD3 | | DD4 | | DD5 | | DD6 | |
|---|---|---|---|---|---|---|---|---|
| | F | Sig. | F | Sig. | F | Sig. | F | Sig. |
| Dragon --> Respeaker | 9.79 | .002* | 9.81 | .002* | 3.16 | .08 | 10.32 | .002* |
| Respeaker --> Corrector | 29.45 | .000* | 28.06 | .000* | 32.19 | .000* | 9.65 | .003* |
| Corrector --> Broadcaster | 7.50 | .008* | 6.17 | .015* | 2.94 | .089 | 2.55 | .114 |

Note: The degrees of freedom (df) is always equal to 1.

As shown in Table 11, the differences between each of the phases are significant in DD3. In other words, the increase in accuracy rate is significant between Dragon and the respeaker, between the respeaker and the corrector, and between the corrector and the broadcaster. The same can be said for DD4. For the last two episodes, the situation is slightly different. In DD5, the difference in accuracy rate between Dragon and the respeaker is not significant, which means that the respeaker was not able to make sufficient corrections and was therefore unable to increase the accuracy rate enough to generate a significant increase in the accuracy rate at this stage. In the next stage, the increase is significant: the corrections made by the corrector increase the accuracy rate significantly. However, in the last stage, there is no further significant increase: the corrections made by the broadcaster do not increase the accuracy rate in a significant way. In the last episode, DD6, there is a significant increase in the accuracy rate from phase 1 to phase 2, that is, between Dragon and the respeaker and between the respeaker and the corrector. However, in the last phase, there is no further significant increase, just as in episode 5.

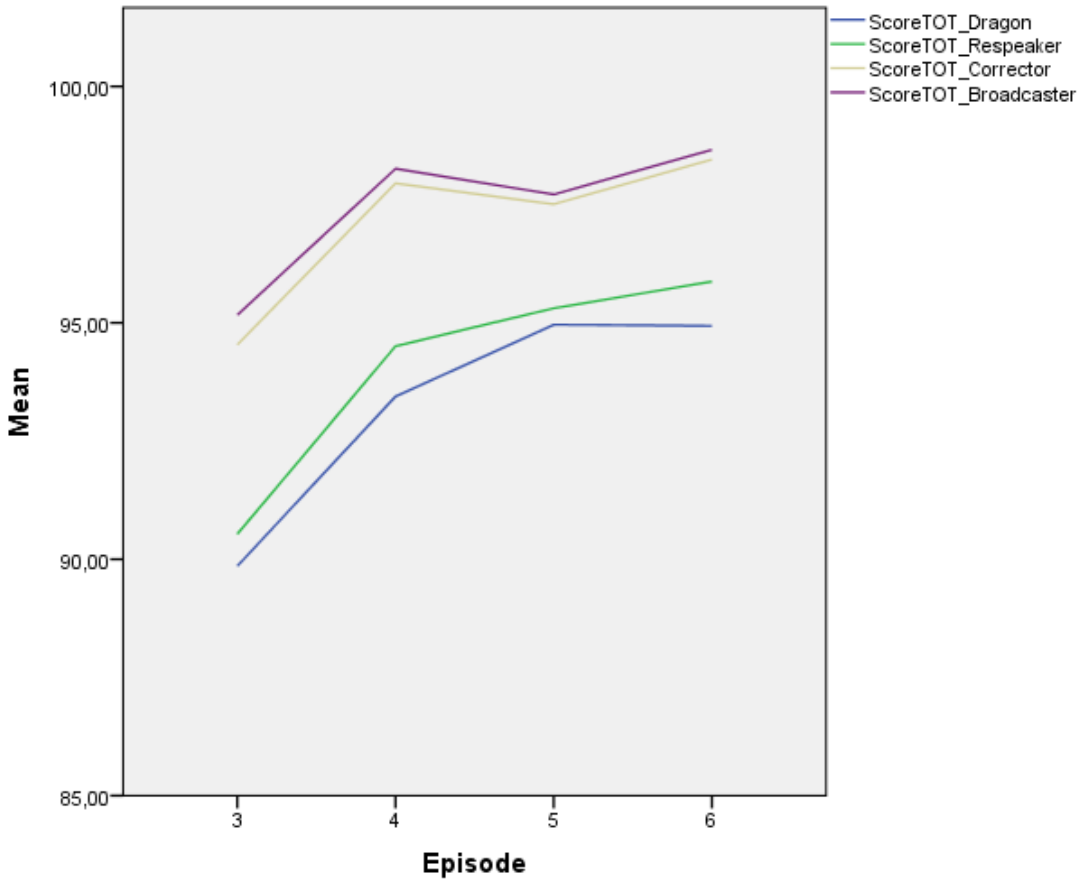


Figure 2: Accuracy rate per member of the team for all episodes

What these results show is that in interlingual live subtitling, the corrector seems to play a significant role in increasing the accuracy rate. For each episode, the corrector improves the accuracy rate significantly. This is, however, what can be expected from that function. The respeaker also manages to increase the rate in some cases. In DD5, there is no significant increase in the accuracy rate, but this is also the episode where the score for Dragon – that is, what Dragon produces on the basis of the respoken words of the respeaker – is the highest. In other words, this respeaker might have concentrated more on respeaking than on correcting and leaving much of this task to the corrector. The last finding is this: in two of the four episodes, the broadcaster does not increase the accuracy rate of the subtitles. The broadcaster's main task being the broadcasting of the subtitles, one cannot expect a significant increase at this stage. The question is therefore whether this task could be automated or reallocated to the STI, who works on one and the same computer with the corrector.

Figure 2 represents the accuracy rate of each team member (ScoreTOT in the legend) across the four episodes. As Table 9 has shown, there is an increase in the accuracy rate in each phase, for each episode. In addition, the increase seems highest between the interventions of the respeaker and the corrector (second and third line, starting from the bottom). What Figure 2 also reveals is that accuracy rates seem to increase for each team member from one episode to another: all the lines are ascending, except for the corrector and the broadcaster, who seem to perform less well in DD5 than in DD4 and DD6. All in all, all the team members seem to have performed better as from DD4, with DD3 starting from the lowest position. However, the question is whether these increases are significant. To answer this question, we carried out the non-parametrical Kruskal-Wallis test for independent groups, since we are now comparing different people: as shown in section 4.2.3, the same

function was almost never fulfilled by the same person. Since the results were all significant, we also conducted the parametrical variant of the test (one-way ANOVA) to be able to examine differences at pair-level (post hoc Bonferroni). The results of both tests are summarized in Table 12.

Table 12: Kruskal-Wallis and ANOVA tests (comparison of several independent groups)

| Kruskal-Wallis | Dragon | R/I | Corrector | Broadcaster |
|---|---|---|---|---|
| Chi-Square | 24.33 | 27.25 | 17.79 | 15.18 |
| Df | 3 | 3 | 3 | 3 |
| Monte Carlo Sig. | .000 | .000 | .000 | .000 |
| One-way ANOVA | | | | |
| F | 6.70 | 7.11 | 7.59 | 6.39 |
| Df | 3 | 3 | 3 | 3 |
| Sig. | .000 | .000 | .000 | .000 |

The results of the post hoc tests (see Appendix 1) revealed significant differences for all the functions between DD3 and DD4, DD3 and DD5, and DD3 and DD6. However, there was no significant difference for the other pairs, that is, DD4-DD5, DD4-DD6 and DD5-DD6. This means that the accuracy rate of Dragon in DD3 was significantly lower than in all the later episodes, but that the differences in the accuracy rate of Dragon in the later episodes were not significant. Consequently, it can be said that Dragon underperformed in DD3, meaning that the respeaker in DD3 respoke significantly less accurately than in the others, since the Dragon accuracy rate is dependent on the respeaker's input. Similarly, the figures show that the respeaker in DD3 performed significantly less well than the respeakers in DD4, DD5 and DD6, but that there is no significant difference in the accuracy rate between the respeakers of DD4, DD5 and DD6. The same applies to the correctors and the broadcasters. What can be concluded at this stage, therefore, is that the very first step of the process – that is, the respeaking and its level of accuracy – is of paramount importance to all subsequent steps. However, it must be borne in mind that the tasks were slightly different each time and that the lower results for DD3 might be due to the particular characteristics of the episode.

### 5.2. Error analysis

Figure 3 represents an overview of all the errors (RE and EE) categorized in section 3. The numbers include all the errors that were made, whether they were corrected at a later stage or not. Figure 4 shows only the remaining errors, that is, those broadcast. A comparison of both figures clearly shows that REs, which are the most prominent type, are efficiently corrected. Small omissions, in contrast, generally remain uncorrected – which is predictable, because they are content-related and therefore mostly dependent on the respeaker. That is, because the corrector does not hear the original audio, they cannot correct omissions, since they are not aware of them. It would normally fall to the STI to detect these, since they do have access to the ST audio. However, this rarely happens, which means that the corrector, even with the assistance of the STI, is not able to intervene in this type of content error or omission. Nevertheless, small omissions are only minor EEs, which means that their impact on user comprehension is limited. They might, however, have an impact on the quality perception of the viewers, if they know enough English, which is usually the case in Flanders. In other words, this may be a topic for follow-up research on reception.

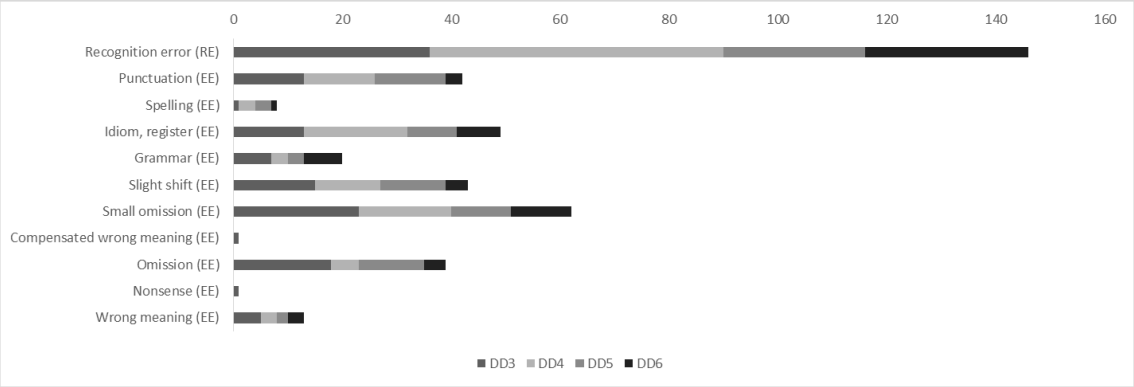


Figure 3: Number of errors made, per type

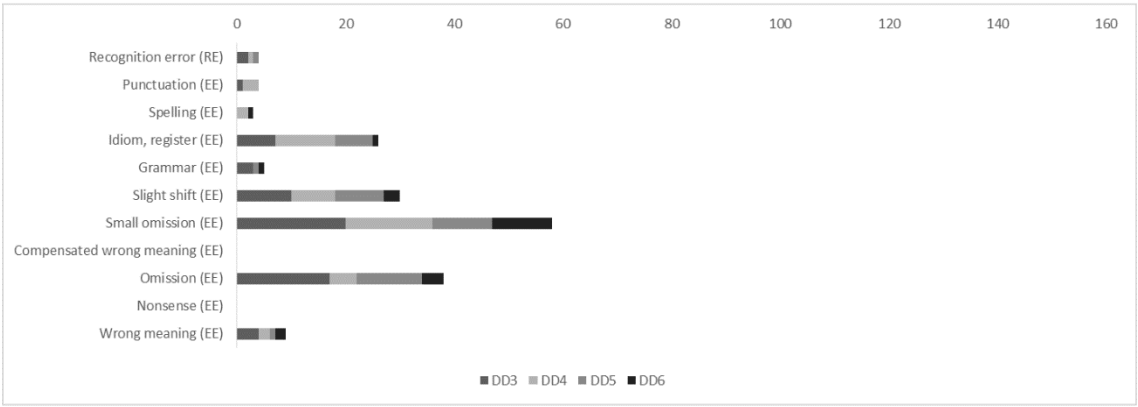


Figure 4: Number of errors remaining in the broadcast episodes, per type

Another way of looking at these results is to determine the correction rate of each member of the team. In other words, this involves identifying the stage at which errors are corrected and determining whether there is a difference in the types of error that are made and corrected, that is, in the correction rate for REs versus EEs.

In Table 13 we provide an overview of the correction rate of each team member, related to the number of remaining errors with which each has to cope rather than the total number of errors made from the beginning.

Table 13: Overview of the correction rate per team member

| Episode | Correction rate respeaker (%) | Correction rate corrector (%) | Correction rate broadcaster (%) | Broadcast errors (%) |
|---|---|---|---|---|
| RE | | | | |
| DD3 | 20.0 | 85.7 | 50.0 | 5.7 |
| DD4 | 23.1 | 92.5 | 66.7 | 1.9 |
| DD5 | 16.7 | 95.0 | 0.0 | 4.2 |
| DD6 | 38.5 | 93.8 | No errors left | 0.0 |
| Mean | 24.6 | 91.7 | 38.9 | 3.0 |
| EE | | | | |
| DD3 | 2.06 | 27.37 | 8.70 | 64.95 |
| DD4 | 6.67 | 28.57 | 6.00 | 62.67 |
| DD5 | 1.54 | 31.25 | 6.82 | 63.08 |
| DD6 | 2.38 | 39.02 | 8.00 | 54.76 |
| Mean | 3.16 | 31.55 | 7.38 | 61.36 |

Table 13 demonstrates that almost all REs are corrected as only an average 3% of REs remain uncorrected and are therefore broadcast. In this process the respeaker corrects on average almost 25% of them (24.6%), the corrector then corrects more than 90% of the remaining RE (91.7%) and the broadcaster manages to correct almost 40% of the remaining errors after the corrections have been made by the corrector (38.9%).

The situation is totally different for EEs, however: almost 62% of them (61.36%) remain uncorrected and are broadcast. Besides that, though, the correction rate is lower for each function or phase. The respeaker does not correct many of the EEs (only 3.16%), since they are also the individuals who make almost all of them. The corrector performs much better, correcting almost 32% of the remaining EEs (31.55%), but their correction rate for EEs is much lower than that for REs. Finally, the broadcaster manages to correct 7.38% of the remaining EEs, which is also much less than their correction rate for REs.

It is important to note that although almost 62% of EEs remain uncorrected, a majority of them (70%) are minor errors, as Figure 5 illustrates. Nevertheless, 22% of EEs are of the standard+ type, a category that has been added to the NER model and whose weight (0.75) is higher than that of standard errors in that model (normally 0.5). This might be the reason why the accuracy rate remains substandard in two out the four *Dansdate* episodes.

As far as the EE category is concerned (Figure 6), the majority of the minor EEs that end up in the broadcast (73%) have an effect on comprehension, either because they are small omissions (48%) or slight shifts (25%). The remaining minor EEs (27%) would probably have an impact on quality perception. Standard EEs all have an impact on perception (n = 5), standard+ EEs are all omissions with loss (missing clause; n = 39), and all serious EEs belong to the comprehension category (wrong meaning, n = 9).

This error analysis therefore shows again that the very first step of the process – that is, respeaking – is of paramount importance. The correction rates of REs are high and do not leave much room for improvement. The correction rates of EEs are much lower but depend even more on respeaking performance: most of the omissions can only rarely be corrected by the corrector or the broadcaster. In other words, this is an area where there is still much room for improvement in the current division of tasks.

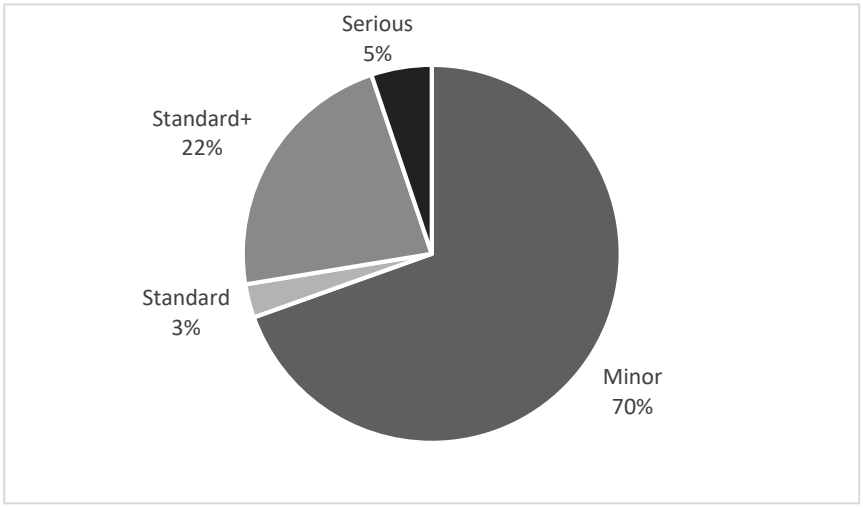


Figure 5: Seriousness level of the broadcast EEs

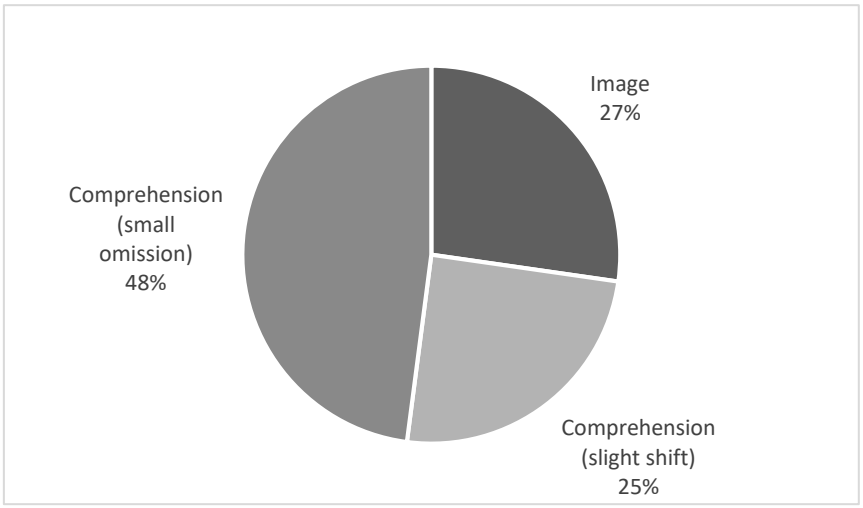


Figure 6: Broadcast minor EEs, per type

### 5.3. Relation between number of errors, accuracy rate and episode features

In this last section, we try to find some explanations for the differences in the accuracy rates between the different episodes. Table 14 summarizes some key features of all four episodes.

Table 14: Key features of all four episodes

| | DD3 | DD4 | DD5 | DD6 |
|---|---|---|---|---|
| Number of words in English | 998 | 1221 | 1164 | 963 |
| Number of respoken words | 552 | 913 | 762 | 681 |
| Number of words in broadcast episode | 542 | 916 | 763 | 672 |
| Average respeaking rate (number respoken words/number of English words) | 63.33% | 81.47% | 73.59% | 79.36% |
| Average number of words in English per corresponding subtitle | 10.98 | 10.71 | 10.49 | 10.58 |
| Average number of respoken words per corresponding subtitle | 6.13 | 8.01 | 6.36 | 7.48 |
| Average number of words in broadcast subtitle | 6.02 | 8.04 | 6.87 | 7.38 |

Note: The higher the respeaking rate, the more verbatim the respeaker is respeaking.

We focus on the following correlations: (1) between the number of words in English and the number of all errors of each type that were made (whether corrected or not); (2) between the number of words in English and the accuracy rate of each function; (3) between the respeaking rate and the number of all errors of each type that were made (whether corrected or not), and (4) between the respeaking rate and the accuracy rate of each function. The results are presented in Appendix 2. Whether all the episodes are taken together or separately, many correlations appear to be significant. However, we will discuss those correlations that are significant to all the episodes taken together and compare them to each episode taken separately.

First, there is a significant *positive* correlation between the number of words in English and the number of REs, EEs and even CEs. This positive correlation applies to all the episodes taken together as far as EEs and CEs are concerned. However, the correlation with the number of REs is significant only in DD3. In other words, in general, the more words that have to be respoken/interpreted, the more the errors that are made.

Second – and this is a rather logical consequence – there is a significant *negative* correlation between the number of words in English and the accuracy rate of all the members of the team. This is true for all the episodes taken together and also for DD5. For DD3, DD4 and DD6 the correlation is significant only with the accuracy rate of the corrector and the broadcaster. In other words, the more that English words have to be respoken and subtitled, the lower the accuracy rate of the corrector and the broadcaster (and sometimes of Dragon and the respeaker).

Third, there is a significant *positive* correlation between the respeaking rate (percentage of respoken words in English in Appendix 2) and the number of REs, but a *negative* correlation between the respeaking rate and the number of EEs. In other words, the more verbatim the respeaker respeaks, the more the REs that are made, but the fewer the EEs that are made. This is true for all the episodes taken together and also for DD3. In DD6, these correlations are not significant. In DD4 and DD5, the correlation is significant (negatively) only with the number of EEs.

Fourth, there is a significant *positive* correlation between the respeaking rate and the accuracy rate of all the members. This is true for all the episodes taken together and also for DD3 and DD5. For DD4, the correlation is not significant with the accuracy rate of Dragon and for DD6, it is not significant with the accuracy rate of Dragon and of the respeaker. However, in general, it can be said that the higher the respeaking rate, that is, the more verbatim the respeaker respeaks, the higher all the members of the team score, especially the corrector and the broadcaster.

Finally, we calculated the correlation rate between the number of words in English and the respeaking rate, on the one hand, and of the remaining (and therefore broadcast) REs and EEs, on the

other hand. No correlation was found with the number of remaining REs, which was to be expected since that number is low, as indicated before. In contrast, there is a significant *positive* correlation between the number of words in English and the number of remaining EEs ($rs = 0.421$, $p<.000$), and a significant *negative* correlation between the respeaking rate and the number of remaining EEs ($rs = -0.299$, $p<0.000$). In other words, the more words there are to respeak, the more the EEs that remain; but the more verbatim the respeaker respeaks, the fewer the EEs that are broadcast. This is true for all episodes taken together or separately.

## 6. Concluding thoughts

Designing a quality-assessment model involving the classification of translation errors that occur in live interlingual subtitling and assigning a weight to the different types of error is an ambitious undertaking and some degree of subjectivity seems unavoidable. We are fully aware of the fact that the present attempt needs to be further adapted, tested and validated, but we are equally convinced that the present study offers a very good starting point. In addition, it yields interesting insights for further research.

Having surveyed some of the literature on quality parameters in subtitling and SI, we proceeded to design a translation-assessment control model for live interlingual subtitling, adapting the well-tested NER model designed for intralingual live subtitling. We then proceeded to test the model with our case study, using the data from four episodes of *Dansdate*, and we concluded with a first assessment of the live interlingual subtitling produced through the procedure currently used at VTM. In doing so, we not only assessed the accuracy of the subtitles as broadcast but also the accuracy of all the different intermediary stages and the performance of the different professionals involved. Quite a few interesting findings emerged.

First of all, our results confirm earlier findings (Van Waes et al., 2013) for live intralingual subtitling in the sense that the delivery rate of the audio has an impact on the performance of the respeaker team, as can be expected. The higher the number of words to be respoken/interpreted, the higher the number of REs and in particular of EEs and, therefore, the lower the accuracy rate (that is, the quality of the output) of all the members of the team. This was not true for all the episodes, but it seems to be a trend. However, when only the remaining broadcast errors are taken into account, the number of words to be respoken no longer has an effect on REs (no correlation). The respeaking rate of the respeaker also seems to have an impact on each step of the process: the higher the respeaking rate, the more verbatim the respeaker respeaks, the higher the accuracy rate of the corrector and the broadcaster. However, there is a considerable difference in the correction success rate of EEs versus REs (the latter being much higher) and this difference persists in each phase of the process. Nevertheless, it is also clear that experience plays a role in the performance of all those involved, since the data show that the accuracy rate achieved for each team member increases substantially from one episode to the next for all types of error.

Looking at the individual interventions, the current corrector has the greatest positive impact on the accuracy rate of the subtitles, followed by the respeaker and the broadcaster, whose impact is minimal. However, both the score of the corrector and that of the broadcaster increase when the respeaking rate of the R/I increases.

This, in fact, is an important finding: the performance of the R/I is carried through to the subtitle as broadcast. In other words, the performance of the respeaker is crucial, since none of the subsequent interventions can improve the quality of the subtitles that results from the subtitling process set off by a poor respeaker to the level achieved by a good respeaker. This finding, as well as the significant difference in the general performance achieved with respect to the correction of REs as opposed to EEs for each function, is certainly food for thought – both for further research and for R/I training and software development.

Indeed, our data show that there are significant differences in the performances of the R/Is in our study, with DD3 performing below average. However, the data we gathered about the R/Is' backgrounds do not enable us to explain this difference. The respeaker for DD3 is an interpreter with

nine years of experience, but what exactly do these nine years indicate about the number of hours spent on interlingual subtitling? In addition, whereas the four episodes may belong to the same genre, they were not controlled for speech rate or degree of difficulty. The lower rating of DD3 is largely due to major omissions, which may have been the result of more overlapping speech. In brief, too many variables are at play for us to explain the difference in performance at this stage. In a subsequent study, the quantitative data will therefore have to be combined with qualitative data, for example those from in-depth interviews. Furthermore, more variables related to the audio input will have to be controlled.

In addition, the classification of the translation errors itself requires further fine-tuning and testing. Whereas, generally speaking, the NER model appears to work for interlingual live subtitling, translation "proper" does present additional challenges. The difference between minor omission errors (EEs) and minor omissions classified as CEs is not always clear-cut in interlingual live. Examples include verbs of opinion in the main clause (of the type "I think that …") that were not respoken and that might often not be included in interlingual pre-prepared subtitling either but that do add nuances of meaning. Other examples are the omission of coordinating conjunctions, also typical of interlingual subtitling, and generalisations (e.g., the use of hypernyms). In this respect interlingual, which involves translation "proper", may differ from intralingual live, which is more verbatim across the board and therefore presents fewer problems in distinguishing minor omissions that count as EEs from minor omissions that count as CEs.

Furthermore, our error analysis and classification may be language-pair bound. The model certainly needs to be tested for different languages and with a larger number of raters. Right now, it is not entirely clear, for instance, why the subtitling performance was only "acceptable" in *Dansdate* with no more than one episode reaching a good score higher than 98 per cent correctness according to the NER model. Since interlingual live subtitling is more challenging than intralingual live, the weighting of errors may have to be adapted. Perhaps the standard+ error rating that we introduced should be abandoned (although it designated major omissions) or perhaps the 98 per cent NER standard is simply too high for interlingual live. Unfortunately, this would seem to imply that a lower standard is acceptable for interlingual live subtitling, which is not desirable. However, this can be assessed only through reception research – another avenue for further research.

The current set-up at VTM seems to be delivering reasonable-quality subtitles but there is room for improvement. The present study suggests that it works well for the correction of REs but performs poorly when it comes to the correction of EEs, especially in the form of unwanted (minor) omissions, which cannot be detected. The data also show that it is important to ensure that the first link in the subtitling chain, the R/I, is carefully selected, whereas the importance of the broadcaster in error detection and correction is minimal. Further research into the live subtitling process may therefore want to investigate alternative set-ups as well as the advantage of designating the different tasks to particular professionals with specific skills (as also suggested by Szarkowska, Krejtz, Dutka & Pilipczuk, 2016), all of which would subsequently have to be fed into further research on subtitler profiles and training. To be continued.

## References

Carroll, M., & Ivarsson, J. (1998). *Code of good subtitling practice*. Berlin: European Association for Studies in Screen Translation.

Díaz Cintas, J., & Remael, A. (2007). *Audiovisual translation: subtitling.* Manchester: St. Jerome.

Kuo, S.-Y. (2014). *Quality in subtitling: Theory and professional reality* (Doctotal dissertation). Retrieved from https://spiral.imperial.ac.uk/bitstream/10044/1/24171/1/Kuo-SzuYu-2014-PhD-Thesis.pdf.

Kurz, I. (2001). Conference interpreting: Quality in the ears of the user. *Meta, 46*(2), 394 –409.

Pereira, A. (2010). Criteria for elaborating subtitles for deaf and hard of hearing adults in Spain: Description of a case study. In A. Matamala & P. Orero (Eds), *Listening to subtitles: Subtitles for the Deaf and Hard of Hearing* (pp. 87–102). Bern: Peter Lang.

Pöchhacker, F. (2016). *Introducing Interpreting Studies.*(2nd ed.). London: Routledge.

Rajendran, D.J., Duchowski, A.T., Orero, P., Martínez, J., & Romero-Fresco, P. (2013). Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives: Studies in Translatology, 21*(1), 5–21.

Remael, A., Van Waes, L., & Leijten, M. (2014). Live subtitling with speech recognition: How to pinpoint the challenges? In D. Abend-David (Ed.), *Media and translation: An interdisciplinary approach* (pp. 121–147). New York, NY: Bloomsbury Academic.

Robert, I.S., & Remael, A. (2016). Quality control in the subtitling industry: an exploratory survey study. *Meta, 61*(3), 578–605.

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

Romero-Fresco, P. (2016). Accessing communication: The quality of live subtitling in the UK. *Language and Communication, 48*, 56–69.

Saldanha, G., & O'Brien, S. (2013). *Research methodologies in translation studies.* Manchester: St. Jerome.

Szarkowska, A., Krejtz, K., Dutka, Ł., & Pilipczuk, O. (2016). Cognitive load in intralingual and interlingual respeaking – a preliminary study. *Poznan Studies in Contemporary Linguistics, 52*(2), 209–233.

Van Waes, L., Leijten, M., & Remael, A. (2013). Live subtitling with speech recognition: causes and consequences of text reduction. *Across Languages and Cultures, 14*(1), 15–46.

Zwischenberger, C. (2010). Quality criteria in simultaneous interpreting: An international vs. a national view. *The Interpreters' Newsletter, 15*, 127–142.

**Appendix 1 (Post hoc tests section 5.1)**

| Dragon | Mean Difference (I-J) | Std. Error | Sig. |
|---|---|---|---|
| DD3–DD4 | −3.58737* | 1.25334 | .027 |
| DD3–DD5 | −5.10432* | 1.26814 | .000 |
| DD3–DD6 | −5.08097* | 1.30943 | .001 |
| DD4–DD5 | −1.51695 | 1.15368 | 1.000 |
| DD4–DD6 | −1.49360 | 1.19892 | 1.000 |
| DD5–DD6 | .02335 | 1.21438 | 1.000 |
| Respeaker | | | |
| DD3–DD4 | −3.97455* | 1.21495 | .007 |
| DD3–DD5 | −4.77658* | 1.22929 | .001 |
| DD3–DD6 | −5.34168* | 1.26932 | .000 |
| DD4–DD5 | −.80203 | 1.11834 | 1.000 |
| DD4–DD6 | −1.36713 | 1,16219 | 1,000 |
| DD5–DD6 | −.56510 | 1.17718 | 1,000 |
| Corrector | | | |
| DD3–DD4 | −3,41249* | ,85719 | ,000 |
| DD3–DD5 | −2.97093* | .86731 | .004 |
| DD3–DD6 | −3.91671* | .89554 | .000 |
| DD4–DD5 | .44156 | .78902 | 1.000 |
| DD4–DD6 | −.50421 | .81996 | 1.000 |
| DD5–DD6 | −.94577 | .83054 | 1.000 |
| Broadcaster | | | |
| DD3–DD4 | −3.09488* | .83581 | .001 |
| DD3–DD5 | −2.55004* | .84568 | .016 |
| DD3–DD6 | −3.49516* | .87322 | .000 |
| DD4–DD5 | .54484 | .76935 | 1.000 |
| DD4–DD6 | −.40028 | .79952 | 1.000 |
| DD5–DD6 | .94513 | .80983 | 1.000 |

**Appendix 2**

| All episodes | | | ScoreTOT_Dragon | ScoreTOT_Respeaker | ScoreTOT_Corrector | ScoreTOT_Broadcaster | Nb_RE | Nb_EE | Nb_CE |
|---|---|---|---|---|---|---|---|---|---|
| Spearman's rho | Nb_Words_Eng | Correlation Coefficient | −.126** | −.162** | −.366** | −.373** | .108* | .413** | .411** |
| | | Sig. (1-tailed) | .007 | .001 | .000 | .000 | .015 | .000 | .000 |
| | Nb_Words_Resp | Correlation Coefficient | .078 | .039 | −.139** | −.127** | .183** | .217** | .278** |
| | | Sig. (1-tailed) | .064 | .222 | .003 | .007 | .000 | .000 | .000 |
| | Nb_Words_Final | Correlation Coefficient | .087* | .051 | −.111* | −.101* | .161** | .202** | .298** |
| | | Sig. (1-tailed) | .046 | .163 | .015 | .024 | .001 | .000 | .000 |
| | Pct_RespokenWords_English | Correlation Coefficient | .213** | .226** | .340** | .368** | .117** | −.240** | −.311** |
| | | Sig. (1-tailed) | .000 | .000 | .000 | .000 | .009 | .000 | .000 |
| DD3 | | | | | | | | | |
| Spearman's rho | Nb_Words_Eng | Correlation Coefficient | −.168 | −.160 | −.462** | −.509** | .223* | .702** | .329** |
| | | Sig. (1-tailed) | .072 | .082 | .000 | .000 | .017 | .000 | .001 |
| | Nb_Words_Resp | Correlation Coefficient | .094 | .133 | −.142 | −.142 | .395** | .458** | .357** |
| | | Sig. (1-tailed) | .207 | .124 | .109 | .109 | .000 | .000 | .000 |
| | Nb_Words_Final | Correlation Coefficient | .117 | .149 | −.082 | −.077 | .347** | .419** | .390** |
| | | Sig. (1-tailed) | .156 | .099 | .238 | .252 | .000 | .000 | .000 |
| | Pct_RespokenWords_English | Correlation Coefficient | .320** | .340** | .550** | .635** | .225* | −.273** | −.111 |
| | | Sig. (1-tailed) | .002 | .001 | .000 | .000 | .017 | .005 | .149 |
| | | N | 77 | 77 | 77 | 77 | 89 | 89 | 89 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| DD4 | | | | | | | | | |
| Spearman's rho | Nb_Words_Eng | Correlation Coefficient | .043 | −.072 | −.379** | −.361** | .073 | .338** | .382** |
| | | Sig. (1-tailed) | .329 | .226 | .000 | .000 | .220 | .000 | .000 |
| | Nb_Words_Resp | Correlation Coefficient | .227** | .115 | −.159* | −.141 | .084 | .183* | .326** |
| | | Sig. (1-tailed) | .008 | .115 | .048 | .070 | .186 | .026 | .000 |
| | Nb_Words_Final | Correlation Coefficient | .248** | .130 | −.127 | −.111 | .046 | .140 | .358** |
| | | Sig. (1-tailed) | .005 | .088 | .094 | .123 | .314 | .069 | .000 |
| | Pct_RespokenWords_English | Correlation Coefficient | .150 | .191* | .309** | .298** | .094 | −.214* | −.249** |
| | | Sig. (1-tailed) | .059 | .023 | .001 | .001 | .162 | .012 | .004 |
| DD5 | | | | | | | | | |
| Spearman's rho | Nb_Words_Eng | Correlation Coefficient | −.267** | −.252** | −.349** | −.330** | .033 | .433** | .438** |
| | | Sig. (1-tailed) | .003 | .005 | .000 | .000 | .366 | .000 | .000 |
| | Nb_Words_Resp | Correlation Coefficient | .019 | .025 | −.089 | −.065 | .100 | .200* | .018 |
| | | Sig. (1-tailed) | .425 | .402 | .184 | .255 | .148 | .017 | .426 |
| | Nb_Words_Final | Correlation Coefficient | −.019 | −.011 | −.106 | -.083 | .113 | .231** | .062 |
| | | Sig. (1-tailed) | .425 | .455 | .142 | .202 | .119 | .007 | .257 |
| | Pct_RespokenWords_English | Correlation Coefficient | .296** | .287** | .319** | .315** | .117 | −.244** | −.535** |
| | | Sig. (1-tailed) | .001 | .002 | .000 | .001 | .110 | .005 | .000 |

| DD6 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Spearman's rho | Nb_Words_Eng | Correlation Coefficient | −.111 | −.143 | −.261** | −.271** | .104 | .208* | .499** |
| | | Sig. (1-tailed) | .150 | .090 | .006 | .005 | .163 | .024 | .000 |
| | Nb_Words_Resp | Correlation Coefficient | −.017 | −.101 | −.156 | −.126 | .103 | .144 | .366** |
| | | Sig. (1-tailed) | .437 | .173 | .071 | .118 | .166 | .086 | .000 |
| | Nb_Words_Final | Correlation Coefficient | −.001 | −.064 | −.132 | −.114 | .099 | .139 | .352** |
| | | Sig. (1-tailed) | .497 | .275 | .107 | .142 | .176 | .095 | .000 |
| | Pct_RespokenWords_English | Correlation Coefficient | .113 | .084 | .234* | .300** | .025 | −.122 | −.367** |
| | | Sig. (1-tailed) | .144 | .214 | .013 | .002 | .406 | .125 | .000 |

1 Verbatim subtitles aim to render the dialogues virtually unabbreviated, editing out only minor conversational features, whereas edited subtitles rephrase and/or delete secondary information to varying degrees.

2 Problems related to subtitle font have so far been related to the limited capabilities of Teletext (or Ceefax), but with the advent of digital television this is gradually being abolished across Europe and beyond.

3 The authors realize that this is conjectural to some extent and that, at a later stage, this tentative typology will have to be tested with users.

4 We have opted for the term “corrector” rather than “editor” because of the variety of errors that are in effect “corrected” by this person. We feel that the corrector does more than “edit” the text. Actually, we feel that the terminology that is now used to refer to the various stages of the live interlingual subtitling process and the professionals involved in it may be in need of a serious overhaul. However, that is beyond the scope of the present article.

5 For the sake of completenes, we would like to point out that VTM feels that perfectly synchronous interlingual live subtitles are not “credible” and therefore require the broadcaster to put the subtitles on air with a very slight delay in order to maintain the live effect for the viewers.

6 *So You Think You Can Dance*, is a franchise of reality television shows in which contestants compete in dance. The first series of the franchise, created by producers Simon Fuller and Nigel Lythgoe, premiered in the United States in July 2005 (https://en.wikipedia.org/wiki/So_You_Think_You_Can_Dance).

7 After each session, all the team members discuss their performance and give one another feedback.