# The use of artificial intelligence in the assessment of live subtitling quality: the NER Buddy

Pablo Romero-Fresco
Óscar Alonso Amigo
Luis Alonso Bacigalupe

**Pablo Romero-Fresco**
Department of Translation and Linguistics, Universidade de Vigo;
promero@uvigo.gal;
ORCID: 0000-0003-2166-5792

**Óscar Alonso Amigo**
Software developer and expert in Artificial Intelligence;
alamios@protonmail.com;
ORCID: 0009-0000-4480-5845

**Luis Alonso Bacigalupe**
Department of Translation and Linguistics, Universidade de Vigo;
lalonso@uvigo.gal;
ORCID: 0000-0002-3020-4536

## Abstract

Translation quality assessment has been subject to high levels of subjectivity. However, in areas such as audiovisual translation it has become common practice to objectively evaluate the quality of the captions of live TV broadcasts. In intralingual live subtitling — an accessibility service for people with hearing loss where captions are in the same language as the original — the NER model was proposed by Romero-Fresco and Martínez (2015). However, it is complex and time-consuming. The purpose of this contribution is to present the results of our research on the development of an AI-based application for the (semi-)automatic assessment of live captions using the NER methodology. International TV broadcasters are testing this app.

Keywords: live subtitling, respeaking, automatic speech recognition (ASR), the NER model, artificial intelligence (AI), large language models (LLMs), automatic assessment, NER Buddy.

## Resumen

La evaluación de la calidad de la traducción siempre está sujeta a altos niveles de subjetividad. Sin embargo, en áreas como la traducción audiovisual, se ha convertido en una práctica común evaluar objetivamente la calidad de los subtítulos de las transmisiones en vivo en televisión. En el subtitulado en vivo intralingüístico —un servicio de accesibilidad para personas con pérdida auditiva en el que los subtítulos están en el mismo idioma que el original—, el modelo NER fue propuesto por Romero-Fresco y Martínez (2015). No obstante, este modelo es complejo y requiere mucho tiempo. El propósito de este artículo es presentar los resultados de nuestra investigación sobre el desarrollo de una aplicación basada en inteligencia artificial para la evaluación (semi)automática de subtítulos en vivo utilizando la metodología NER. Actualmente, varias cadenas internacionales están probando esta aplicación.

Palabras clave: subtitulación en vivo, respeaking, reconocimiento automático del habla (ASR), modelo NER, inteligencia artificial (IA), modelos masivos de lenguaje (LLM), evaluación automática, NER Buddy.

*Rebuda: 20 de novembre de 2023 | Acceptació: 4 de novembre de 2024 | Publicació avançada: 20 de desembre de 2024*

Revista Tradumàtica. Número 22 (Desembre 2024)
**Citació recomanada:** ROMERO-FRESCO, PABLO; ALONSO AMIGO, ÓSCAR; ALONSO BACIGALUPE, LUIS (2024). «The use of artificial intelligence in the assessment of live subtitling quality: the NER Buddy». Revista Tradumàtica. Tecnologies de la Traducció, 22, 450-470. <https://doi.org/10.5565/rev/tradumatica.408>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

**Resum**

L'avaluació de la qualitat de la traducció sempre està supeditada a alts nivells de subjectivitat. En canvi, en àrees com ara la traducció audiovisual, s'ha convertit en una pràctica comú avaluar objetivament la qualitat dels subtítols de les transmissions en viu a la televisió. En el subtitulat en viu intralingüístic —un servei d'accessibilitat per a persones amb pèrdua auditiva en què els subtítols estan en el mateix idioma que l'original—, el model NER ha estat proposat per Romero-Fresco i Martínez (2015). No obstant això, aquest model és complex i requereix molt de temps. El propòsit d'aquest article és presentar els resultats de la nostra recerca sobre el desenvolupament d'una aplicació basada en inteligència artificial per a l'avaluació (semi)automàtica de subtítols en viu utilitzant la metodologia NER. Actualment, vàries cadenes internacionals estan provant aquesta aplicació.

**Paraules clau**: subtitulació en viu, respeaking, reconeixement automàtic de la parla (ASR), model NER, intel·ligència artificial (IA),models massius de llenguatge (LLM), avaluació automàtica, NER Buddy

## 1. Background on live subtitling

Intralingual live subtitling (ILS) is an accessibility service where live subtitles (or captions, the preferred term in the United States (US) and Australia) in the same language as the audio are shown on the screen for people with hearing loss and for the wider hearing audience wishing to use them. Live captions are becoming increasingly popular these days. Although it is estimated that in the US, 14% of Americans have a hearing loss (HLAA 2018), subtitles for the deaf and hard-of-hearing are used regularly by 50% of Americans, which increases to 70% in the case of members of Generation Z (Mykhalevych 2022).

Live subtitles initially broadcast in the UK and the US in the early 1980s were produced using keyboards, either standard QWERTY keyboards, dual keyboards (with a team of two live subtitlers working on the same programme), or special keyboards such as the Velotype, which allows the user to press several keys simultaneously and produces syllables rather than letters (Lambourne 2006). This method was soon replaced by stenography, where the subtitler can press multiple keys at the same time to spell out not only whole syllables and words, but also phrases, with a single hand motion and much faster than with the Velotype. Stenography, however, is a complex and expensive technique that requires extensive training (between three and four years), more than the time required to operate the Velotype (Marsh 2006).

In the early 2010s, a new method known as respeaking was introduced as an alternative to the previous workflows (Romero-Fresco and Eugeni 2020). In intralingual respeaking, "a respeaker listens to the original sound of a (live) programme or event and respeaks it, including punctuation marks and some specific features for the deaf and hard-of-hearing audience, to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay" (Romero-Fresco 2011:1). Respeaking (or speech-to-text-interpreting (STTI)) (Stinson 2015) has now become the most popular live subtitling method all over the world.

However, the development of automatic speech recognition (ASR) software and the introduction of new developments in artificial intelligence (AI) have brought about a new method that is gradually gaining ground: fully automatic live subtitling, which transcribes the original speech and identifies different speakers without the need for any human intervention. Given that the introduction of this live subtitling method in online meeting platforms has increased exponentially after the 2020 worldwide pandemic, broadcasters and companies all over the world are beginning to combine human and fully automatic live subtitling, especially because the latter is considerably more affordable than the former (Romero-Fresco and Fresno 2023).

As the aim of live subtitling is to provide accessibility for users, some of whom (mostly those with hearing loss) would not otherwise be able to access the audiovisual content, comparative research on the quality of these different methods is crucial. The aim of this paper is to present NER Buddy, the first AI-based software that can (semi-) automatically analyse the quality of live subtitles. Following an introduction in Section 2 to the NER model, the most commonly used method to assess live subtitling quality, in Section 3 the article addresses the role of AI in language assessment. Section 4 analyses the development and results of the training process undergone by the new software. This is followed by a discussion on the current limits of AI in the assessment of live subtitling quality (Section 5) and a final reflection on the potential ways forward in this area (Section 6).

## 2. The assessment of live subtitling quality: the NER model

Different methods have been developed over time for the assessment of live subtitling quality. One of the most frequently used ones is the Word Error Rate (WER) model (see Figure 1), which has served as a basis for other methods. In the WER model, N is the total number of words and there are three different types of errors: Substitution (a correct word is replaced by an incorrect one), Insertion (an extra word is added) and Deletion (a correct word is omitted) (Dumouchel et al. 2011).

$$\text{WER} = \frac{S + D + I}{N} \times 100$$
$$\text{Accuracy} = 100 - \text{WER}$$

*Figure 1: The WER model to assess the accuracy of live subtitles*

The problem with this model is that it does not account for different degrees of error severity (Wells et al. 2022). All errors are penalised as -1, regardless of whether or not they have an impact on the users' comprehension of the subtitles. Another issue is that this model does not account for instances in which a subtitler edits, condenses or paraphrases the original text without necessarily changing or losing meaning. As is the case with translation, live subtitling is not about thoughtlessly reproducing every word of the original, but about conveying the intended meaning of the message. This could be illustrated by instances where the subtitles omit unimportant asides or fillers from the transcript (*you know*, *I mean*, *kind of*, *um*), which is a useful strategy commonly applied by subtitlers to keep up with the speech rate of the original speaker, but also to avoid unnecessary "noise" in the subtitles which may hamper efficient communication.

The NER model (Romero-Fresco and Martínez 2015), shown in Figure 2, accounts for different types of errors. The N stands for the number of words in the subtitles. The E stands for edition errors, those stemming from strategies applied by the subtitler (omissions, for example), and the R stands for recognition errors, that is, misrecognitions. These edition and recognition errors are classified as serious (when they introduce misleading but credible information), standard (when they cause confusion and loss of information) or minor (when they do not impact on comprehension), scoring -1, -0.5 and -0.25 respectively. The NER model also identifies correct editions (CEs), which are instances in which the subtitler's editing has not led to a loss of information. In this model, a 98% accuracy rate is the minimum quality threshold required, as originally established following user tests in the DTV4ALL project (Romero-Fresco 2015) and as confirmed in more recent reception studies in Canada and Poland (CRTC 2019a, Romero-Fresco 2020).

$$Accuracy = \frac{N - E - R}{N} \times 100$$

CE (correct editions):
Assessment:

*Figure 2: Formula used by the NER model to calculate accuracy*

The NER model has gradually become the standard for live subtitling quality assessment around the world. It is used by broadcasters, media companies and universities and it is included in the official accessibility guidelines in countries such as Canada, the UK and Spain (CRTC 2019, UNE 2012). However, it is not without problems. In order to focus on meaning rather than words and to account for how different types of errors impact on comprehension, it requires a human evaluator to compare the transcripts of the speech and the subtitles, which means that the model is both time-consuming and subject to some degree of subjectivity. In contrast, WER is fully automatic, as it relies exclusively on word count, but it can also be inaccurate (Wells et al. 2022), as seen above.

Live subtitling companies have long been requesting an automatic tool that can apply the NER model with little or no human intervention, given the financial and human resources that they are regularly devoting to this purpose. NER Buddy, presented in this paper, is the first attempt to develop this automatic tool.

## 3. AI in language processing: the NER Buddy

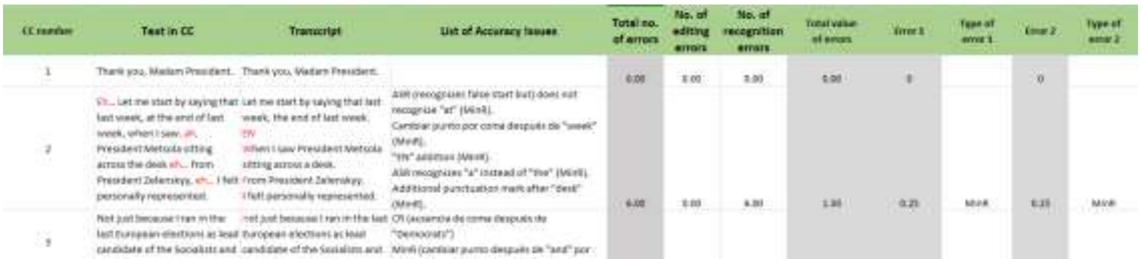
Until recently, developing a system to automatise the NER model (which, unlike WER, is based on meaning, discriminates error severity and involves human-like decision-making) was a pipe dream. The field of language modelling through machine learning and neural networks (usually known as AI) has now made this possible by bringing together the benefits of automaticity (including consistency, speed of analysis and low cost) and the advantages of human evaluation, where decisions about quality are not based on differences in word count, but on the implications that such textual differences might have for effective and successful communication.

Language modelling through AI has a relatively long history, but one of its watershed moments was the recent publication of the milestone paper “Attention is all you need” (Vaswani et al. 2017), which put forward the transformer architecture that finally outperformed the more traditional recurrent neural network architecture. This was followed by the launch in 2018 of Google’s Bert (Devlin et al. 2019), an encoder-only transformer model that is still widely used these days, and of OpenAI’s GPT-1, a decoder-only transformer model. These models are pretrained on large amounts of texts to perform different tasks, such as text classification or sentence prediction. Since then, the field has progressed rapidly with a steady growth in the size of both the model and the corpus used for training. A further milestone was the release of GPT-3 in 2020 (Brown et al. 2020). The model had an unprecedented size of 175 billion parameters, which allowed for the generation of high-quality texts, as well as for performance of certain tasks with only a few examples (few-shot). However, these developments went largely unnoticed by the general public until OpenAI’s release, in November 2022, of GPT-3.5 and its popular interface, ChatGPT, which was capable of performing a number of language-related tasks after receiving instructions in natural language exclusively, with a level of accuracy never seen before. GPT-4, launched by OpenAI in March 2023, is still one of the most popular state-of-the art models. Ever since, although there has been a substantial improvement in the values of the benchmarks used to evaluate the models — for example, MMLU is the benchmark for reading comprehension (Papers with Code 2024) — it is not clear if the capabilities of LLMs to undertake real tasks have improved substantially in terms of handling text. However, new modalities have been added, such as visual and auditory capabilities, and the models have been renamed as large multimodal models (LMMs).

Following these developments, the field flourished rapidly and a number of actors began to compete with each other in a race to push the capabilities of these models to the next level. The launch of LlaMA by META in February 2023, with an open access system for research activities exclusively, was another milestone that paved the way for open-source models. Since then, thousands of open models have been released, from relatively small (but still capable) models that can be run on a personal computer to large models that rival the state-of-the-art commercial models.

AI-based systems are bound to bring about disruptive transformations to the landscape of automatic language processing, including ILS. Just as respeaking replaced stenography in the 2010s as the preferred method all over the world, ASR is now becoming increasingly common, so much so that one of the leading ILS companies worldwide, AI-Media, is now for the first time providing more automatic than human-made subtitles (Ward 2024). Needless to say, a decision with such significant professional implications must be underpinned by solid data regarding subtitling quality, which in the case of AI-Media is based on NER results, hence the importance of having a tool that can assist with the application of this model.

Until now, most NER evaluations all over the world have resorted to tailor-made Excel spreadsheets (see Figure 3). Evaluators compare the transcript (placed in one column) with the subtitles (placed in the next column), identify the discrepancies between them, and decide whether these are correct editions or minor, standard or serious errors.



| CC number | Text in CC | Transcript | List of Accuracy Issues | Total no. of errors | No. of editing errors | No. of recognition errors | Total value of errors | Error 1 | Type of error 1 | Error 2 | Type of error 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Thank you, Madam President. | Thank you, Madam President. | | 0.00 | 0.00 | 0.00 | 0.00 | 0 | | 0 | |
| 2 | Eh... Let me start by saying that last week, at the end of last week, when I saw, ah, President Metsola sitting across the desk eh... from President Zelenskyy, eh... I felt personally represented. | Let me start by saying that last week, the end of last week, TV when I saw President Metsola sitting across a desk. from President Zelenskyy. I felt personally represented. | ASR (recognises false start but) does not recognise "at" (MinR). Cambiar punto por coma después de "week" (MinR). "TV" addition (MinR). ASR recognises "a" instead of "the" (MinR). Additional punctuation mark after "desk" (MinR). | 6.00 | 0.00 | 6.00 | 1.50 | 0.25 | MinR | 0.25 | MinR |
| 3 | Not just because I ran in the last European elections as lead candidate of the Socialists and | not just because I ran in the last European elections as lead candidate of the Socialists and | CR (ausencia de coma después de "Democrats") MinR (cambiar punto después de "and" por | | | | | | | | |

Figure 3: Tailor-made Excel spreadsheet for the manual use of the NER model

Although built-in formulas allow for automatic calculation of accuracy rate, as well as type and number of errors, evaluators still have to transcribe the clip, align transcript and subtitles, and identify the discrepancies before deciding what type of error to score.


Figure 4: Example of automatic analysis with the NER Buddy

NER Buddy is an AI-based application developed in 2023 by members of the GALMA research group, as part of the Spanish Ministry of Education-funded QUALISUB project (https://qualisub.webs.uvigo.es/inicio/), with the aim of (semi-)automating the use of the NER model. The software allows for three types of modes: semi-automatic, automatic with revision, and fully automatic (without revision). In the semi-automatic mode, the software aligns transcript and captions and identifies discrepancies between them. The evaluator selects the types of errors manually and the software calculates accuracy rate and other relevant numerical data. In the automatic mode, once the transcript and the captions have been automatically aligned and the discrepancies between them have been identified, the software applies the NER model, scoring different types of errors and calculating the final accuracy rate. This may or may not be followed by a human revision. Figure 5 provides a comparative estimate of the time it would take to analyse a 10-minute clip with the NER model using a custom-made Excel spreadsheet and using NER Buddy in its semi-automatic mode and automatic modes with and without revision. This estimate, which is pending confirmation in further studies, has been obtained through consultation with five NER evaluators involved in the QUALISUB project and with NER evaluators from the leading international captioning company AI-Media.

| Task | Excel | Semiautomatic NER Buddy | Automatic NER Buddy with revision | Automatic NER Buddy without revision |
| --- | --- | --- | --- | --- |
| **Transcription** | 1h | 1h | 1h | 1h |
| **Alignment** | 1h | 0h | 0h | 0h |
| **Analysis** | 2h | 2h | 1h | 0h |
| **Results** | 1h | 0h | 0h | 0h |
| **Total** | **5h** | **3h** | **2h** | **1h** |

*Figure 5: Comparison of time efficiency in the semi-automatic, automatic with revision, and fully automatic without revision modes vs. manual Excel spreadsheets*

The main contribution that NER Buddy can offer, as compared to the prevailing Excel-based manual approach, is thus automatic alignment and, most importantly, automatic analysis, which draws on the latest AI-based language models (LLMs) and is crucially underpinned by extensive training.

At this point, NER Buddy is in beta phase and is being tested by international actors in the audiovisual industry, as well as by academic and research partners. It is also open, on request, to other stakeholders willing to test it.

## 4. Training LLMs

AI-based LLMs are built on the principle of anticipation of information, i.e. rather than understanding language properly they retrieve solutions from previous "experience" in memory. The term "stochastic parrot" (Bender et al. 2021) has been coined to describe this. Therefore, just like humans, LLMs require intensive training to gain experience in the field of expertise in which they are being instructed, so as to make informed choices for future instances where similar or identical problems might emerge. This is a process whereby massive amounts of data are incorporated into the model so that it recognises frequent sentence patterns and establishes connections between words and sentences in the text, which will be the basis for the subsequent decisions and judgements the AI will have to make.

### *4.1. Pre-training, instruction tuning, prompting and fine-tuning*

During the pre-training stage, LLMs are fed with information packages made up of vast amounts of texts, books and webpages, without instructions, which can be relevant to different types of activities. Upon completion of this phase, the LLM should be able to autocomplete texts and to perform certain simple tasks following a small number of examples that are illustrative of those tasks (few-shot).

For the instruction tuning phase, the LLM is provided with instruction-response pairs (with the instructions in natural language only, i.e. zero-shot) to guide the model on how to solve tasks.

### *4.2. Training the NER Buddy*

In order to help NER Buddy achieve human-like accuracy in the assessment of subtitling quality using the NER model, we put the software through two alternative modes of training: prompting and fine-tuning.

#### *4.2.1. Prompting*

The training stage was performed firstly through prompting. This involves the introduction into the system of sets of detailed instructions which may contain different communicative styles and formats to interact with the chatbot, including questions, conversations, requests for explanations of the rationale followed, examples and clarifications of error categories. The prompts must be adjusted as many times as needed until the automatic system provides consistent answers and avoids significant deviations from the values obtained by human assessment. Minor changes (no matter how small they may be) introduced in the formulation of the sets of instructions used in this phase are likely to have an impact on the answers provided by the system. It is worth noting that, when undergoing this process, the LLM is not learning in a way that enables it to make future inferences, which means that the LLM needs to receive all the instructions and examples whenever a new question is asked. Three different approaches to prompting were tested: few-shot, forked-chain (a customised version of the prior chain-of-thought methodology (Wei et al. 2023), which was not tested in this research; see below), and division into clauses.

During few-shot prompting, two versions of ChatGPT were tested: gpt-3.5-turbo-0301 and gpt-4-0314. They were fed with basic indications on how to use the NER model and with a small number of relevant but simple examples, with no more than one error per line of text to make the learning process easier, as shown in Figure 6. Our work was based on previous training done by Speechmatics, one of the leading actors worldwide on the speech recognition stage.

```
[user]: Work out the severity of errors based on the comparison between the
reference and recognised transcript. Consider the effect on what a human viewer
could understand. The three error types are:
Minor: The sentence is still easy to understand.
Standard: The sentence is nonsensical.
Serious: The sentence appears to make sense, but it changes the meaning. It's
misleading or causes misinformation.

Here are some examples:
Reference="the first topic is the supreme court."
Recognised="the first top sick is the supreme court."
{ "error": "standard", "reason": "'topic' is misrecognised as 'top sick' causing
the sentence to be unclear." }

[...]

Now, try yourself. Responds only with JSON:
Reference="he asked me one simple question."
Recognised="he asked me once a question."

[assistant]: { "error": "serious", "reason": "'simple' is misrecognised as 'once'
 which changes the meaning of the sentence." }
```

*Figure 6: Example of few-shot prompting*

Chain-of-thought prompting is a methodology where an intermediate step is included with the aim of helping the model follow a reasoning process, instead of asking for a direct answer to specific questions. It is usually combined with the few-shot methodology and, rather than testing it for this research, it served as an inspiration for the newly developed forked-chain method described below.

Forked-chain is a new methodology implemented by the team (see Figure 7). First, the LLM is prompted with a chain of intermediate questions intended to expand the amount of information available to inform the final decision. Afterwards, the model is asked to follow a chain of decisions between two options or forks: whenever a choice is made, another one is ruled out, which helps to guide the LLM in the right direction towards a particular rationale and result.

```
[user]: I will give you step-by-step instructions. Answer as briefly as you can.
Compare the following sentences:
Reference="the first topic is the supreme court."
Recognised="the first top sick is the supreme court."

What is the meaning of the Reference sentence?
[assistant]: The Reference sentence means that the initial subject of discussion
is the Supreme Court.

[user]: What is the meaning of the Recognised sentence?
[assistant]: The Recognised sentence is likely a mishearing or typo of the
Reference sentence, intended to mean that the initial subject of discussion is
the Supreme Court.

[user]: Are those meanings equal or equivalent? Explain why.
[assistant]: No, the meanings are not equivalent. The Recognised sentence
contains "top sick" which doesn't make sense in this context. It seems like a
misinterpretation or typo of "topic" from the Reference sentence.

[user]: Considering the above questions, is the difference in meaning between the
 two sentences minor or important? Answer with one word: minor or important.
[assistant]: Important

[user]: Is the meaning of the Recognised sentence clear and makes sense or is it
senseless? Answer with one word: sense or senseless.
[assistant]: Senseless
```

*Figure 7: Example of forked-chain prompting*

Upon initial observation, NER Buddy proved to be often unable to identify more than one agent (subject) or more than one action (verb) in individual lines of subtitles sent for assessment. This meant that a significant number of differences between original transcript and subtitles were going unnoticed by the LLM, impacting on its performance. The team then opted for division into clauses, asking the newly launched GPT-4 about the number of verbs in the subtitles and showing that wherever there is a verb in a sentence there is also normally a subject, and wherever there is a subject plus a verb there tends to be clause, i.e. a unit of propositional meaning that requires grading. GPT-4 was first asked to divide the subtitle into the different clauses (by ascertaining how many actions and actors were included in a given subtitle) and then to analyse each individual clause (see Figure 8). However, this approach was not successful in addressing the challenge of analysing multi-clause and multi-error subtitles, and was abandoned due to lack of reliability.

```
[user]: I will give you step-by-step instructions. Answer as briefly as you
can.
Compare the following sentences:
Reference="in fact, they have been the ones that have been bombing and shelling
 eastern aleppo."
Recognised="in fact, they have been the ones that up in bombing and schelling
eastern aleppo."
Divide both Reference and Recognised sentences into clauses as short as
possible. Use the following JSON format:
{"Clause 1": {"Reference": "...", "Recognised": "..."}, "Clause 2":
{"Reference": "...", "Recognised": "..."}}

[assistant]:
{
    "Clause 1": {
        "Reference": "in fact,",
        "Recognised": "in fact,"
    },
    "Clause 2": {
        "Reference": "they have been the ones",
        "Recognised": "they have been the ones"
    },
    "Clause 3": {
        "Reference": "that have been bombing",
        "Recognised": "that up in bombing"
    },
    "Clause 4": {
        "Reference": "and shelling eastern aleppo.",
        "Recognised": "and schelling eastern aleppo."
    }
}
```

```
[user]: Divide Clause 4 into subjects, verbs and each object separately.
Preserve contractions. Use the following JSON format:
{"Subject 1": {"Reference": "...", "Recognised": "..."}, "Verb 1":
{"Reference": "...", "Recognised": "..."}, "Object 1": {"Reference": "...",
"Recognised": "..."}}

[assistant]:
{
    "Subject 1": {
        "Reference": "",
        "Recognised": ""
    },
    "Verb 1": {
        "Reference": "shelling",
        "Recognised": "schelling"
    },
    "Object 1": {
        "Reference": "eastern aleppo.",
        "Recognised": "eastern aleppo."
    }
}

[user]: In the context of the whole Recognised sentence compared to Reference
sentence, is the diference in meaning of Verb 1 relevant or irrelevant? Answer
with one word: relevant or irrelevant.
[assistant]: Relevant
```

*Figure 8: Example of prompting with division into clauses*

## *4.2.2. Fine-tuning*

The second training mode tested with NER Buddy was customised fine-tuning. This consisted in feeding the LLM with 1940 examples of NER scoring performed by human experts (including transcripts, subtitles and results) and containing a total of 3472 errors (see Annex for further details), so that it could learn on its own by extrapolating solutions (effectively making inferences) from the examples originally provided. All the data used for both training and testing were provided by commercial broadcasters for research purposes and are, therefore, real-life materials.

Also, since ChatGPT is a commercial product, fine-tuning was done on the OpenAI platform. The versions put to the test were, first, gpt-3.5-turbo-0613 and, second, gpt-3.5-turbo-1106. It is hard to work out the amount of time required for preparation of the materials subject to analysis, as this includes both the time spent in designing the data processing pipeline and the time required to carry out the human NER analysis needed to train and test the efficiency of the AI. Although this phase can initially be time-consuming, once the training has been completed a brief instruction is all the software needs to perform a NER analysis. Unlike in the prompting training mode, the LLM does not need to receive all the instructions and examples for every new analysis.

Hyperparameter optimisation was carried out, testing the following parameters: batch size 1, 2 and 8, and learning rate (LR) multiplier 0.5 and 2. For the final set-up, the base model was gpt-3.5-turbo-1106, batch size 2, LR multiplier 2, for two epochs. The fine-tuning process took as much as two hours for a total of 2,852,026 trained tokens, which are the basic units of information that LLMs handle and are equivalent to approximately three quarters of a word.

### *4.3. Results*

In order to test the potential of the first training mode (prompting), we ran a preliminary test with NER Buddy: a selection of 300 examples of NER human assessment (evaluated by members of our research group) was prepared, and as many as 72 sets of prompts were drafted throughout the process to help NER Buddy achieve as much consistency as possible with human evaluation. As can be seen in Table 1 below, the overall level of agreement between human and automatic assessment was 71.43%, that is, 64.29% for minor errors (agreement in 27 cases out of a total of 42), 74.51% for standard errors (38 out of 51) and 83.33% for serious errors (10 out of 12).

| | | Human | | | |
|---|---|---|---|---|---|
| | | Minor | Standard | Serious | Total |
| GPT-3.5 | Minor | **27** | **4** | **1** | **32** |
| | Standard | **12** | **38** | **1** | **51** |
| | Serious | **3** | **9** | **10** | **22** |
| | Total | **42** | **51** | **12** | **105** |

*Table 1: Results of preliminary tests with prompting. Confusion matrix showing human-NER Buddy agreement for each error category. The green boxes highlight matches between GPT and human. The red boxes show discrepancies.*

Once our methodology had been refined and the results appeared to be going in the right direction, a first round of tests with real broadcast materials taken from different TV genres in Spanish was launched. The results from automatic assessment were not, however, too encouraging: NER Buddy appeared to be identifying many more discrepancies between transcripts and subtitles than human evaluators, leading to automatic scores (i.e. accuracy rates) that were considerably lower than those provided in human assessments.

A careful analysis revealed that these results were not necessarily caused by the materials analysed (real TV broadcasts, as opposed to the list of individual examples from prior human NER analyses used in the preliminary tests) but by the technique used to produce the subtitles. Respeakers often edit a great deal of the original content, sometimes condensing (i.e. omitting or paraphrasing) as much as 40% of the original audio, as they need to ensure that the speech recognition software they are dictating to recognises every word they say. This causes problems for NER Buddy, which may be tempted to identify every word missing in the subtitles as an error. A second round of tests was thus launched, this time using ASR-produced subtitles, which are normally near-verbatim — in other words, they attempt to transcribe every word of the original audio. Although only three five-minute samples were tested, the results were more acceptable, with the automatic assessment of one of the genres (news) falling “only” 2.55% short of the human assessment. Still, on average, NER Buddy scores were 6.29% below human assessment (see Table 2) and the system was identifying three times more errors than human evaluators.

| Language | Subtitling tool | NER Human | NER GPT | Gap human/ GPT | Sample size |
|---|---|---|---|---|---|
| ES | ASR subtitles | 98.23% | 91.94% | 6.29 points | 15 min. |

*Table 2: Results of the second round of tests with prompting*

For the final test with prompting — the third round of tests — we used different audiovisual genres (speeches, interviews, news and talk shows) and, more importantly, subtitles produced by state-of-the-art AI-based ASR systems in English (17 minutes) and Spanish (15 minutes): more specifically, we tested Ursa (Hughes 2023) and Whisper Large V2 (Radford et al. 2022). As shown in Table 3, NER Buddy produced much better results, falling only an average of 0.5% short of the results obtained by human evaluators.

| Language | Subtitling tool | NER Human | NER GPT | Gap human/ GPT | Sample size |
|---|---|---|---|---|---|
| ES | AI-based ASR | 98.88 % | 98.37 % | 0.51 points | 17 min. (85 ex.) |
| EN | AI-based ASR | 99.53 % | 99.11 % | 0.42 points | 15 min. (75 ex.) |

*Table 3: Results of the third round of tests with prompting*

It would thus seem that the use of state-of-the-art ASR subtitles (which featured fewer errors than the previous set of ASR subtitles and fewer omissions than respoken subtitles) played a crucial role in this analysis.

However, despite the headway achieved through successive waves of prompt-engineering (step-by-step instructions vs. more general instructions, division into clauses vs. full sentences, identification of principal grammatical categories, etc.), the scores produced by NER Buddy reached a plateau, which opened the door to the testing of a new approach (fine-tuning). Fine-tuning, or the ability of AI to learn on its own when fed with relevant examples, is one of the main benefits of AI and LLMs. Although it is currently believed that a LLM can be fine-tuned with 90-150 examples, we decided to feed gpt-3.5 with 900 examples of NER errors assessed by humans, corresponding to approximately 180 minutes of spoken text. The reason for this is that the task to be undertaken by NER Buddy is not as "simple" as writing a poem or a research paper, which only requires mimicking someone else's text based on previous experience. Instead, NER Buddy is required to make human-like judgements over human language, that is, to decide whether or not meaning (rather than words) is lost in the subtitles with regard to the original transcription of the audio. The results obtained in this case, included in Table 4, were very encouraging, as NER Buddy proved to be only 0.15% off the score produced by human assessment.

| Language | Subtitling tool | Human | GPT | Gap human/ GPT | Sample size |
|---|---|---|---|---|---|
| ES + EN | AI-based ASR | 98.13 % | 98.00 % | 0.13 points | 200 examples (40 min.) |
| ES + EN | AI-based ASR | 98.13 % | 98.18 % | 0.05 points | 200 examples (40 min.) |
| ES + EN | AI-based ASR + STTI | 98.15 % | 98.40 % | 0.25 points | 500 examples (100 min.) |

*Table 4: Results after fine-tuning for specific tasks (or self-learning)*

Nonetheless, a word of caution is required here. While these results suggest that NER Buddy is applying the NER model more proficiently, a much greater sample of real-life materials and real-life tests is required to confirm this trend. International media access companies, such as Apptek and AI-Media, are currently testing the software, which will help to ascertain if it is accurate and efficient enough to merit in-house use in the assessment of live subtitling quality.

## 5. The (current) limits of AI in the assessment of live subtitling quality

While the training undergone so far by NER Buddy has shown very promising results, it has also revealed some aspects that remain challenging and that point more generally to the limits of LLMs as we currently know them. Firstly, NER Buddy has shown a certain tendency towards inconsistency, which has recently been identified as one of the main issues in the use of LLMs for language evaluation. As shown by Stureborg et al. (2024), LLM evaluators often change their judgments for different samples, demonstrating significantly lower inter-sample agreement than human experts' inter-annotator agreement. Indeed, LLMs have occasionally produced different results for two evaluations of the same set of subtitles and, when engaged in a conversation with the instructor, they have sometimes been shown to change their output in an unpredictable manner. An extreme example is instances of hallucinations, that is, factual inconsistencies that can lead to completely fabricated answers, which are currently being studied as a frequent occurrence in the use of LLMs for language evaluation (Tang et al. 2024). Although not often, ChatGPT has sometimes been found to hallucinate, such as when it (momentarily) assigned a completely different set of rules and different authors to the NER model, despite the information it had been fed.

It is important to acknowledge, however, that it is not uncommon for different human evaluators or even the same human evaluator to change their assessment of a specific set of subtitles when revising it. As a matter of fact, one of the main issues encountered in the training of NER Buddy is the inconsistency of the human evaluations fed into the software. This does not necessarily mean that different human evaluators understand the rules of the NER model differently, but rather that they understand live subtitling differently. Those evaluators who consider live subtitling as a form of transcription that enables speech-to-text-based communication do not expect live subtitles to adhere to all the grammatical and syntactical conventions that may be required of standard written language. For them, the absence of a comma after "however" in a sentence such as "However, he decided to go" or the use of a comma instead of a full stop to separate two individual sentences (both of which are likely to occur when subtitles are produced by ASR) would be examples of correct editions. In contrast, other evaluators see live subtitling as written language that is going to be read by viewers who may expect or need subtitles to adhere to grammatical and syntactical conventions so that they can be understood. This includes Deaf viewers whose first language is sign language and who will effectively be reading subtitles in their second or third language, immigrants and other second-language viewers, and elderly people who may struggle to read language that is not properly punctuated. Evaluators at this end of the spectrum would score minor errors for the two commas mentioned above, as, even though the meaning of the subtitles is not affected, the use of commas is incorrect and therefore disruptive for the viewers. At the beginning of 2024, the leading international live subtitling company AI-Media asked the NER Buddy team to work with them on a specific set of criteria that could help to ensure consistency in the use of the NER model. A happy medium was then found between the consideration of live subtitles as a transcription and their need to adhere to the rules of written language. The aim was to distribute these criteria amongst human evaluators in order to improve the consistency of their scores. NER Buddy, which was initially trained on strict subtitles-as-written-language data, will also be fed with these new criteria, which may help reduce the discrepancy between its scores and those of human evaluators.

Another aspect that has been challenging in the use of LLMs for language evaluation is their sensitivity, especially regarding the way in which they are prompted (Pezeshkpour et al. 2023). ChatGPT proved to be sensitive to even minor changes in the instructions provided, which makes it essential to consider carefully what words are used in prompting. A case in point is the difficulties ChatGPT had in distinguishing between the different degrees of error severity established by NER (minor, standard and serious). Minor errors did not pose a significant problem, as it proved easy for the AI to ascertain when minor information changes or losses had occurred. However, the system seemed unable to distinguish between standard and serious errors. After several failed attempts to fix this by, for example, changing the way the questions were formulated, it became clear that the problem was not in the formulation of the questions, but in the names of the categories. Everything fell into place as soon as the categories were renamed with more descriptive names that ChatGPT could understand and make sense of: "standard error" was replaced by "nonsensical information" and "serious error" was replaced by "misleading information". This was a watershed moment in our research and a crucial step forward in our results, and it shows the importance of finding the right language to speak to LLMs so that they can perform efficiently.

Finally, another aspect with which we had to struggle is context, especially when it comes to multimodal content. A case in point can be seen in Figure 9, which shows a primary school teacher explaining how sound works to her students in class. At some point, the teacher says:

> "Now, I'm not a good drawer but I'm going to try, right?
> Because we still try things that we are not so good at."

The ASR engine subtitled this as:

> "Now, I'm not a good draw but I'm going to try, right?
> Because we still try things that we're not so good at."

NER Buddy scored a serious error for the misrecognition of "drawer" as "draw" in the subtitles. Strictly speaking, the AI is right, as a new meaning has been created in the subtitles and it can be misleading for users. However, the image (a teacher holding a pen in front of a whiteboard) and the linguistic context (she has just said "don't laugh at my picture") makes it clear that she has said "drawer" instead of "draw", which means that this would be a minor error, that is, one that does not impact on users' comprehension. For NER Buddy to score this correctly, it would need to take into account both the linguistic and visual context of the scene, which, at least until now, has not been possible, but will probably become feasible in the near future when the LMMs mentioned above are launched.


*Figure 9: First example of the use of multimodal context in the use of the NER model*

Another example showing the importance of (visual) context may be found in the live subtitles produced for sports programmes. When subtitling live matches, respeakers are often instructed not to subtitle play-by-play descriptions, that is, the commentators' speech that refers to what is being seen on screen (e.g. "James passes the ball to Hayden"). The rationale is that subtitling this content would prevent viewers from seeing the on-screen action, as they would be reading a description of it which would most likely be shown a few seconds after the event, given the inevitable delay at play in live subtitles. Human NER evaluators are aware of this and normally score instances in which subtitles omit the play-by-play description found in the original audio as correct editions. NER Buddy, however, struggles to understand this, as shown in Figure 10, where an 87-word play-by-play description uttered by a commentator has been omitted in the subtitles and is scored by the software as a case of multiple errors.

*Figure 10: Second example of the use of multimodal context in the use of the NER model*

## 6. Ways forward

The data analysed in the preparation of this paper point to different conclusions and ways forward. One of them inevitably relates to the production of live subtitles and, more specifically, the role of human subtitlers. Live subtitles have so far been produced using different methods, including velotyping, stenography and respeaking, which is still the most commonly used approach all over the world. Given the rapid development of ASR and its increasing accuracy, which in some contexts now rivals human accuracy, it is hard to see how ASR is not going to take over soon as the method preferred by broadcasters and access providers, as shown by the above-mentioned acknowledgement by the leading ILS company AI-Media that, for the first time, they are providing more automatic than human-made live subtitles (Ward 2024). This transition is bound to have a negative impact on live subtitlers. However, although a full analysis of this new scenario exceeds the scope of this paper, it is worth highlighting that human-made live subtitles are still likely to be needed in many contexts, namely those in which ASR cannot produce enough accuracy (for instance, due to the presence of noise, overlapping speech, accents, etc.) or in which it is important to provide non-verbatim or edited content, eliminate or revise errors, etc. In other words, human live subtitlers are likely to be needed when access must be nursed.

Regarding the (semi-)automatic assessment of live subtitling quality, the results obtained with NER Buddy show that it is possible to train state-of-the-art AI-based LLMs to apply an evaluation method such as the NER model. The main goal is to ensure that the scores produced by NER Buddy fall within ±0.4% of the scores obtained by human evaluators (the internal aim set by the team) and as close as possible to ±0.1%, which is the inter-evaluator discrepancy obtained in the largest NER evaluation conducted so far worldwide. Although there is room for improvement, the latest results, shown in Table 4 above, seem promising.

Another interesting way forward concerns the use of different languages in the assessment of live subtitling quality. NER Buddy has been trained in English and Spanish, as shown by the data reported in this article. Yet, recent anecdotal tests in French and Basque, two completely new languages for the software, have yielded accurate results within ±1% of human evaluation scores. Although further investigation is needed, this suggests that a significant part of the learning process undergone by NER Buddy has happened in a language-independent manner, which bodes well for the use of the software in the assessment of languages other than English and Spanish. This also begs the question of whether software such as NER Buddy can conceivably be used in the assessment of interlingual live subtitles, which would require the use of the NTR model (Romero-Fresco and Pöchhacker 2017), that is, the adaptation of the NER model for the evaluation of live subtitles that involve translation (e.g. Spanish audio translated live into English subtitles). Given the positive results obtained recently in the use of LLMs for the evaluation of translation quality (Knocmi and Federman 2023), this seems like a realistic goal to pursue in the near future.

Lastly, it is worth adding a final word about the role of human evaluators in the (semi-)automatic assessment of live subtitling quality. Although NER Buddy is still being rolled out to different live subtitling companies around the world and there are no results available regarding its efficiency, human evaluators are reacting very positively to the amount of time that the software saves in the most tedious tasks, such as alignment and error detection. Indeed, rather than replacing human evaluators, the main goal of NER Buddy is to assist them so that they can use their skills in the areas of language assessment that truly require human judgement.

More generally, the experience obtained until now in the development of NER Buddy has shown us that an endeavour such as this one requires truly interdisciplinary collaboration between software developers, AI experts and linguists. This indicates yet another way in which the skills of translators and interpreters will be required in the era of generative AI.

## References

Bender, E. M., Gebru, T., MacMillan-Major, A. and Schmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp 610–623. <https://doi.org/10.1145/3442188.3445922>. [Accessed: 20240825].

Brown, T. et al. (2020). Language Models are Few-Shot Learners. <https://doi.org/10.48550/arXiv.2005.14165>. [Accessed: 20240815].

CRTC: Broadcasting Notice of Consultation CRTC 2019-9. Ottawa. <https://crtc.gc.ca/eng/archive/2019/2019-9.htm>. (2019a). [Accessed 20240625].

CRTC: Broadcasting Regulatory Policy CRTC 2019-308. Ottawa. <https://crtc.gc.ca/eng/archive/2019/2019-308.htm>. (2019b). [Accessed 20240625].

Devlin, J., Chang, M., Kenton, L., Toutanova, K. (2019). ]BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://doi.org/10.48550/arXiv.1810.04805>. [Accessed 20240621].

Dumouchel, P., Boulianne, G. and Brousseau, J. (2011). Measures for quality of closed captioning, in: A. şerban, A. Matamala and J. M. Lavaur (eds.). *Audiovisual translation in closeup: Practical and theoretical approaches*. Bern: Peter Lang, pp. 161–172.

HLAA (Hearing Loss Association of America): *Hearing Loss: Facts and Statistics* (2018). <https://www.hearingloss.org/wpcontent/uploads/HLAA_HearingLoss_Facts_Statistics.pdf?pdf=FactStats>. [Accessed 20240511].

Hughes, J., (2023). Introducing Ursa from Speechmatics. Speechmatics. <https://www.speechmatics.com/company/articles-and-news/introducing-ursa-the-worlds-most-accurate-speech-to-text>. [Accessed 20240908].

Kocmi, T. and Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality, in *European Association for Machine Translation* (EAMT). <https://arxiv.org/abs/2302.14520>. [Accessed 20240807].

Lambourne, A. (2006). Subtitle Respeaking, in Carlo Eugeni and Gabriele Mack (eds.). *Intralinea, Special Issue on Respeaking*. <https://www.intralinea.org/specials/article/1686>. [Accessed 20240906].

Marsh, A. (2006). Respeaking for the BBC, in Carlo Eugeni and Gabriele Mack (eds.). *Intralinea, Special Issue on Respeaking*. <https://www.intralinea.org/specials/article/Respeaking_for_the_BBC>. [Accessed 20240906].

Mykhalevych, N. (2022). Survey: Why America is obsessed with subtitles, <https://preply.com/en/blog/americas-subtitles-use/>. [Accessed 20222010].

Papers with Code (2024). Multi-task Language Understanding on MMLU. <https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu>. [Accessed 20240908].

Pezeshkpour, P. and Hruschka, E. (2023). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. <https://doi.org/10.48550/arXiv.2308.11483>. [Accessed 20240525].

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. Open AI. <https://cdn.openai.com/research-

covers/language-unsupervised/language_understanding_paper.pdf>. [Accessed 20240908].

Radford, A., Wook Kim, J., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. <https://doi.org/10.48550/arXiv.2212.04356>. [Accessed 20240910].

Romero-Fresco, P. (2011). *Subtitling Through Speech Recognition: Respeaking*. Routledge: Manchester.

Romero-Fresco, P. (2020). Negotiating quality assessment in media accessibility: the case of live subtitling. *Universal Access in the Information Society* 20, pp. 741–751. <https://doi.org/10.1007/s10209-020-00735-6>. [Accessed 20240602].

Romero-Fresco, P. and Martínez, J. (2015). Accuracy rate in live subtitling: the NER model, in Díaz-Cintas, J., Baños, R. (eds.). *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape*. London: Palgrave MacMillan, pp. 28–50. <https://doi.org/10.1057/9781137552891_3>. [Accessed 20240602].

Romero-Fresco, P. and Eugeni, C. (2020). Live subtitling through respeaking, in Bogucki, Ł. and Deckert, M. (eds.). *Handbook of Audiovisual Translation and Media Accessibility*. London: Palgrave MacMillan, pp. 269–297. <https://doi.org/10.1007/978-3-030-42105-2_14>. [Accessed 20240602]..

Romero-Fresco, P., & Fresno, N. (2023). The accuracy of automatic and human live captions in English. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, *22*. <https://doi.org/10.52034/lans-tts.v22i.774>. [Accessed 20240715].

Stinson, M. S. (2015). Speech-to-text interpreting, in Pöchhacker, F. (ed.), *Routledge Encyclopedia of Interpreting Studies.* Manchester: Routledge, pp. 399-40.

Stureborg, R., Alikaniotis, D. and Suhara, Y. (2024). Large Language Models are Inconsistent and Biased Evaluators. <https://doi.org/10.48550/arXiv.2405.01724>. [Accessed 20240908].

Tang, L., Shalyminov, I., Wing-mei Wong, A., Burnsky, J., Vincent, J.W., Yang, Y., Singh, S., Feng, S., Song, H., Su, H., Sun, L., Zhang, Y., Mansour, S. and McKeown, K. (2024). TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization. <https://paperswithcode.com/paper/tofueval-evaluating-hallucinations-of-llms-on>. [Accessed 20240525].

UNE (2012). *Subtitulado para personas sordas y personas con discapacidad auditiva*. Madrid: UNE. <https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0049426>. [Accessed 20240521].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017). Attention Is All You Need. <https://doi.org/10.48550/arXiv.1706.03762>. [Accessed 20240502].

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. <https://doi.org/10.48550/arXiv.2201.11903>. [Accessed 20240716].

Wells, T., Christoffels, D., Vogler, C., Kushalnagar, R. (2022). Comparing the Accuracy of ACE and WER Caption Metrics When Applied to Live Television Captioning, in Miesenberger, K., Kouroupetroglou, G., Mavrou, K., Manduchi, R., Covarrubias Rodriguez, M., Penáz, P. (eds.). *Computers Helping People with Special Needs*. ICCHP-AAATE 2022. Lecture Notes in Computer Science, vol 13341. Springer, Cham. <https://doi.org/10.1007/978-3-031-08648-9_61>. [Accessed 20240602].

## Annex: Results of customised fine-tuning with gpt-3.5-turbo-1106: training data, testing data and statistics

| TRAINING DATA | | | | |
|---|---|---|---|---|
| SAMPLES | 3130 | | | |
| SAMPLES WITH ERROR | 1940 | | | |
| TYPE OF SUBTITLES | ES/ASR | ES/STTI | EN/ASR | |
| | 1108 | 405 | 427 | |
| ERROR SEVERITY | CORRECT | MINOR | STANDARD | SERIOUS |
| ES | 1447 | 1065 | 250 | 37 |
| EN | 277 | 338 | 42 | 16 |

| TESTING DATA | | | | |
|---|---|---|---|---|
| SAMPLES | 709 | | | |
| SAMPLES WITH ERROR | 489 | | | |
| TYPE OF SUBTITLES | ES/ASR | ES/STTI | EN/ASR | |
| | 278 | 102 | 109 | |
| ERROR SEVERITY | CORRECT | MINOR | STANDARD | SERIOUS |
| ES | 386 | 268 | 57 | 9 |
| EN | 55 | 89 | 11 | 4 |

| STATISTICS: NUMBER OF CHARACTERS PER SUBTITLE | | | | |
|---|---|---|---|---|
| | AVERAGE | MEDIAN | Q1 | Q3 |
| ALL | 54.0004 | 43 | 34 | 69 |
| ES | 57.6455 | 45 | 36 | 73 |
| EN | 41.1269 | 35.5 | 25 | 55 |