Accuracy Rate in Live Subtitling – the NER Model

Pablo Romero-Fresco
Juan Martínez

Abstract
For some time now, subtitling companies have been providing broadcasters with regular data on the accuracy of their live subtitles. In some cases this is actually a contractual obligation, as companies are required to obtain a given accuracy rate in their subtitles. However, given that accuracy calculations vary greatly between countries and even companies, the question arises of whether we are effectively comparing incomparable data. The aim of this paper is, first of all, to review different accuracy models used around the world. Then, using the model put forward in Romero-Fresco (2011) as a starting point, a new model is presented to assess the accuracy of live subtitles in different countries and in different languages by analyzing the extent to which errors affect the coherence of the subtitled text or modify its content. Real-life examples in English, Spanish, Italian and German obtained from a corpus of 35,000 words are included. The focus is placed on respoken subtitles, which are nowadays the most common type of live subtitles. However, the model is also applicable to automatic subtitles, since, given the rapid progress of speech recognition technology, they are likely to be introduced in the near future. The model is currently being used by regulators, broadcasters, companies and training institutions in Spain, France, Italy, Switzerland, Germany, Belgium and Australia, among other countries.

1. Introduction

Over the past few years, Audiovisual Translation (AVT) seems to have shifted its focus from quantity to quality. As shown by international conferences such as Media for All 3 (2009)[1] and Media for All 4 (2011)[2], this shift applies both to the industry and to academia. In the case of live subtitling, and more specifically respeaking, the most common method to assess the quality of subtitles produced in real time is usually to assess their accuracy or lack thereof. Needless to say, quality issues in this area involve other features of these subtitles, such as delay, position, character identification and speed, as well as factors that are related to their reception by the viewers (opinion, comprehension, perception), which in the case of the UK have been reported in Romero-Fresco (2011). Yet, what concerns broadcasters, regulators such as Ofcom and subtitling companies is the accuracy of live subtitles, which is what the present paper will deal with.

Up until now, subtitling companies have tackled this issue in very different ways. In some companies the trainers are in charge of error calculation, whereas in others this is done by the subtitlers. The calculation methods vary greatly, some being much more "generous" than others. Furthermore, live subtitling is approached differently depending on the country (see 3.2.3). Whereas in the UK, live subtitles are nearly verbatim, other countries such as Germany or Switzerland produce live subtitles containing different degrees of editing. This heterogeneous picture prompts several questions: Are the accuracy rates obtained by different companies comparable at all? Do the methods used account for the differences between languages? Do they only provide a figure or do they also give an indication of what needs to be improved so as to obtain better results?

The aim of this paper is to present a new model to assess the accuracy of live subtitles in different countries and in different languages by analyzing the extent to which errors affect the coherence of the subtitled text or modify its content. The focus will be placed on respoken subtitles, which are nowadays the most common type of live subtitles. However, the model is also applicable to automatic subtitles, since, given the rapid progress of speech recognition technology, they are likely to be introduced in the near future.[3]

Following an introduction outlining the basic requirements that such a model may be expected to fulfill, an overview is given of the traditional methods used in what is known as word error rate (WER). The next section presents the NER model, including examples in English, Spanish, Italian and German. Finally, an application of the NER model to real-life subtitles will be presented.

2. Basic requirements and traditional methods

[1] www.mediaforall.eu/all3

[2] www.imperial.ac.uk/humanities/translationgroup/mediaforall4

[3] In automatic subtitling, a speaker-independent speech recognition engine (commonly known as ASR or automatic speech recognition) transcribes what the speaker is saying without the need for a respeaker to act as a middle person. This transcription may be shown a) directly as subtitles on screen with no delay (re-synchronization) and no correction, as is the case in a pilot project conducted by the Portuguese television RTP or b) by means of an operator who edits the transcription live (reviewing possible misrecognitions or errors of punctuation and character identification) before launching the subtitles on air with little delay, as done by the Japanese broadcaster NHK.

Before presenting the NER model, it is important to introduce a series of basic requirements that a model such as this may be expected to fulfil so that it can be functional both in academia and in the industry:

2.1. Basic requirements

- Be functional and easy to apply: although the use of multiple variables may be helpful for the researcher, respeakers and trainers must be able to apply the model on a daily basis.[4]

- Include the basic principles of WER calculations in speech recognition theory, which have been tested and validated for long.

- Take into account that different programmes may entail/require different degrees of editing. Sports, for example, are often heavily edited whereas the subtitles for the news, especially in the UK, are near verbatim (Eugeni 2009).

- Take into account that the approach to live subtitling may differ from country to country, thus allowing for the possibility of edited (summarized, expanded, etc.) and yet accurate respeaking. This is the reason why it is not possible to automatise the assessment of accuracy in live subtitling, as has been done in the US (Apone et al. 2010);[5] at least not when respeaking is used.

- Compare the original spoken text with the respoken subtitles to ascertain if there are edition or recognition errors, which may be classified as serious, standard or minor depending on how they affect the processing of textual and visual information.

- Include, whenever possible, other relevant information regarding live subtitling quality, such as delay, position, speed, character identification, etc.

- Provide not only an idea of the quality in terms of accuracy but also of what must be improved (and perhaps even how). Instead of a spot-the-error exercise, the model should provide food for thought as far as training is concerned.

2.2. Traditional WER methods

The US National Institute of Standards and Technology distinguishes between word correctness and word accuracy, both of which are presented as percentages using the following basic formula:

$$\text{Accuracy rate} = \frac{N - \text{Errors}}{N} \times 100 = \%$$

In this model, N is the total number of words spoken by the user. As illustrated by Dumouchel et al. (2011) in the following example, there are at least three different types of errors that can occur with the use of speech recognition: deletion (a correct word is omitted in the recognized sentence), substitution (a correct word is replaced by an incorrect one) and insertion (an extra word is added).

Where	*is*	*the*	*whole*	*wheat*		*flour*
		D	S	S	I	S
Where	*is*		*hole*	*we*	*eat*	*flower*

[4] At the time of writing this article, SWISS TXT, a subsidiary of the Swiss public broadcasting corporation (SRG SSR idée suisse), has commissioned CaiaC (Centre for Research in Ambient Intelligence and Accessibility in Catalonia) with the development of a tool that allows a quick and effective application of the NER model to live subtitles produced by respeaking or by automatic speech recognition (ASR). The tool compares the timecoded transcription of an original spoken text with the timecoded subtitle file by automatically highlighting the differences between the two texts and by colour-coding the different types of errors. The tool works semi-automatically in the background and includes the possibility of manually allocating different values ([1], [0.5] or [0.25]) to different types of errors (serious, standard or minor) as well as incorporating an overall assessment. It also provides statistical data regarding reading speed and delay of the live subtitles analysed by comparing the timecoded information of the transcription and the subtitle file. Subtitle position and character identification may also be analysed, in this case using the original video.

[5] Most live subtitles in the US are produced by steno, with very little editing, which allows for a completely automatic comparison between the source text and the target text.

Taking this into account, the measure of word correctness proposed by the US National Institute of Standards and Technology, which includes deletion and substitution errors, would apply to the above utterance as follows:

$$\text{Accuracy rate} = \frac{N - D - S}{N} \times 100 = 33\%$$

The model to assess word accuracy is stricter, as it also factors in insertion errors:

$$\text{Accuracy rate} = \frac{N - D - S - I}{N} \times 100 = 16\%$$

Designed as they are for the use of speech recognition, these models pose a significant problem when applied to respeaking, as they do not account for instances in which a respeaker edits the original text without changing or losing meaning:

Well, you know, you have to try and put out a good performance, I mean, yeah, it's kind of a stepping stone, isn't it, really?	You have to try to put out a good performance. It's a stepping stone.

$$\text{Accuracy Rate} = \frac{25 - 11 - 1 - 0}{25} \times 100 = 52\%$$

The example above features the omission of relatively unimportant asides (*you know, I mean, kind of*), which constitutes a useful strategy commonly applied by respeakers to catch their breath and keep up with the original speaker. Traditional WER methods would yield an accuracy rate of 52%, whereas a model suited to respeaking may consider this respoken subtitle as 100% accurate.

2.3. The CRIM method

One of the first attempts to adapt traditional WER methods to the specificity of respeaking was carried out by the Centre de Recherche Informatique de Montréal (CRIM, www.crim.ca/en/r-d/reconnaissance_parole). The basis is still the word accuracy method described above, but a step is added in between: once the spoken and the respoken text have been automatically aligned, a human operator goes through the text and decides whether or not the deletions have caused loss of information. In this way, both verbatim and edited respeaking can be accounted for.

Yet, a number of problems remain unsolved. First of all, the decision of when a deletion brings about loss of information is entirely subjective and may thus vary from person to person. This issue will be dealt with in the next section. Secondly, while requirements 1 to 4 above are met, 5 and 6 are not. The accuracy rate obtained with this model may provide useful data, but the deletion figure is ambiguous. Indeed, the formula does not give any indication as to whether the deletions have been caused by misrecognitions or by poor editing strategies on the part of the respeaker. This is a very important distinction, as it requires two different remedial actions. If the deletion errors are mostly misrecognitions, further work is needed on the software to improve the voice profile by fine-tuning the acoustic and language models. In contrast, if the deletions are caused by poor editing by the respeaker, the training should be based on providing the respeaker with skills to edit the original text without losing (excessive) information. Finally, this method does not seem to take into account the specificities of different languages or the occurrence of errors that are subsequently corrected on air by the respeaker. With regard to the latter, some companies, such as IMS in the UK, consider them as half an error, while others do not regard them as mistakes at all, which obviously yields better overall accuracy rates.

3. The NER model

The present model is based on the NERD model included in Romero-Fresco (2011: 150-161). It has been adapted to suit the needs of different types of live subtitling, particularly respeaking and automatic subtitling. The following distinction between serious, standard and minor errors is based on the findings regarding viewers' preferences obtained in the EU-funded DTV4ALL project (www.psp-dtv4all.org), and acknowledges that not all errors pose the same problems in terms of comprehension, thus highlighting the viewer-centred nature of this model.

3.1. Main components of the model

Live subtitles may be expected to reach 98% accuracy using the following model:

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

CE (correct editions):
Assessment:

N: Number of words in the respoken text, including commands (punctuation marks, speaker identification etc.) and words.

E: Edition errors, usually caused by the strategies applied by the respeaker. In other words, they are the result of the respeaker's judgment or decision. The most common situation is that, in a given instance, and for whatever reason (for example because the original speech rate is too fast), the respeaker decides to omit something, thus losing an idea unit (a piece of information)[6]. It could also be that the respeaker adds idea units or paraphrases the original text losing information or introducing wrong information, which could be due to an error in the comprehension of the original text. Editing errors are calculated by comparing the respoken text and the original text and may be classified as serious, standard or minor, scoring 1, 0.5 and 0.25, respectively (specific examples are included in the next section). In the case of automatic subtitles, edition errors are, among others, those related to incorrect capitalization, punctuation and speaker identification.

R: Recognition errors; they are usually misrecognitions caused by mispronunciations/mishearing, or those caused by the specific technology used to produce the subtitles. These errors may be insertions, deletions or substitutions, and are calculated by comparing the respoken text and the original text. They may be classified as serious, standard or minor, scoring 1, 0.5 and 0.25 respectively (specific examples are included in the next section).

CE: Correct editions, that is, instances in which the respeaker's editing has not led to a loss of information, which is calculated by comparing the respoken text and the original text. Given the difficulty involved in producing verbatim live subtitles, the omission of redundancies and hesitations may be considered as cases of correct edition and therefore not as errors as long as the coherence and cohesion of the original discourse are maintained.

Assessment: this section includes the assessment and analysis of the results, as well as comments on different issues, such as the speed and delay of the subtitles, how the respeaker has coped with the original speech rate, the overall flow of the subtitles on screen, speaker identification, the audiovisual coherence between the original image/sound and the subtitles, whether or not too much time has been lost in the corrections, etc. Given that it is difficult to describe the overall quality of a set of live subtitles in a single figure, it is this assessment and not the accuracy rate that determines the quality of subtitles in the NER model.

3.2. Types of errors in English, Spanish, Italian and German

This section includes examples of types of errors obtained from a corpus of 35,000 words made up of live subtitles from some of the most watched news programmes in the UK, Spain, Italy and Switzerland: BBC Six O'Clock News and SKY News (English), 59 segundos and Telediario on TVE (Spanish), Telegiornale RAI (Italian) and 10vor10 and Tagesschau on SF1 (German).

[6] According to Chafe (1985:106), idea units are "units of intonational and semantic closure". They can be identified because they are spoken with a single coherent intonation contour, preceded and followed by some kind of hesitation, made up of one verb phrase along with whatever noun, prepositional or adverb phrase are appropriate, and usually consist of seven words and take about two seconds to produce.

As mentioned above, the approaches to live subtitling may vary from country to country and even from programme to programme. Subtitling companies or broadcasters will often provide clear indications of what is expected from respeakers in terms of editing, summarizing, etc. (see 3.2.3).

3.2.1. Serious errors

Serious errors change the meaning of the original text, creating a new meaning which could make sense in that particular context. Serious recognition errors are often caused by substitutions such as *alms* instead of *arms* or *15%* instead of *50%*. Serious edition errors are often caused by bad choices or confusion on the part of the respeaker (an error with figures or numbers, a change from a positive to a negative statement, etc.). From the viewers' point of view, serious errors do not only omit information but also misinform, which is why some deaf viewers refer to them as "lies". More worryingly, since these errors make sense in the particular context in which they occur, they are usually not noticed by the viewers.

Here are some examples of serious recognition errors in different languages:

> English: *he's having problems with the cheques* instead of *he's having problems with the Czechs*, *they allow young people to smoke but only outside* instead of *they allow young people to smoke pot only outside*, *he never talks to Rudy* instead of *he never talks dirty*, *he was born in 1986* instead of *he was born in 1996*, *the driver must have had a view* instead of *the driver must have had a few.*
>
> Spanish: *casas habitadas por humanos* instead of *casas habitadas por rumanos*, *esto se llama asimetría* instead of *esto se llama simetría*, *siempre usa mal esa conjunción* instead of *siempre usa mal esa conjugación*, *normalmente nos dan ejemplos* instead of *normalmente no se dan ejemplos*.
>
> Italian: *una richiesta di fiducia che nasce* instead of *una richiesta di sfiducia che nasce*, *per tutte le forze armate* instead of *per tutte le forze alleate*, *enorme* instead of *le norme*, *se ci dicono anche che il fidanzato del PD* instead of *se ci dicono anche chi è il fidanzato del PD*, *governo internazionale* instead of *governo di unità nazionale.*
>
> German: *in Island ist der Zinssatz von 3,5 auf 2% gesunken* instead of *in Irland ist der Zinssatz von 3,5 auf 2% gesunken*, *die Schwangerschaft ist ein Zustand, eine Krankheit, man sollte einfach normal weiterleben* instead of *die Schwangerschaft ist ein Zustand, keine Krankheit, man sollte einfach normal weiterleben.*

The following are examples of serious edition errors, which, as in the case of the previous examples, create a new meaning that could make sense in that particular context:

> English:
> *There are a number of questions in the UK, but the big one is whether it's about to slip into recession*.
> instead of
> *There are a number of questions in the UK, but the big one is the US and whether it's about to slip into recession*.
> Spanish:
> *En el interior hay problemas. Ahí sí que se necesita ayuda.*
> instead of
> *En el interior hay problemas, pero no tantos como en la costa. Ahí sí que se necesita ayuda.*
> Italian:
> *dopo la nomina è però arrivata una nota sull'opportunità política*
> instead of
> *dopo la nomina è però arrivata una nota del Quirinale in cui si esprime una riserva del Presidente sull'opportunità politica*
> German:
> *es wird nützen, dass die USA bekannt gemacht haben, dass sie ihre Regierungscomputer besser schützen lassen will*
> instead of
> *es wird wenig nützen, dass die USA bekannt gemacht haben, dass sie ihre Regierungscomputer besser schützen lassen will*

3.2.2. Standard errors

Despite not creating a new meaning, standard errors result in the omission of an information unit from the original text. Standard recognition errors disrupt the flow/meaning of the original text and often cause surprise. They are identified as errors but it is not always easy to figure out what was originally meant (*hell of even* instead of *Halloween*).[7] The difference between standard edition errors and minor edition errors is based on the distinction between independent and dependent idea units. An independent idea unit, such as "The blaze started this morning at the front of the house", is the oral equivalent of a sentence, makes sense as a full, independent message and may be composed of several dependent idea units, such as "this morning" and "at the front of the house". A dependent idea unit is often a complement and it provides information about the "when", the "where", the "how", etc. of an independent idea unit. Standard edition errors often consist in the omission of a full independent idea unit (which may not be noticed by the viewers) or of a dependent idea unit that renders the remaining unit meaningless or nonsensical, i.e. the omission of "last night" in "The rain started last night".

Here are some examples of standard recognition errors in different languages:

English:
Way man Republican instead of *Weimar Republic*, *paid in full by pizza* instead of *paid in full by Visa*, *he's a rats public and* instead of *he's a Republican*, *he's a buy you a bull asset* instead of *he's a valuable asset, attend Tatian* instead of *a temptation* and *I couldn't hear Iran said* instead of *I couldn't hear your answer*.
Spanish:
los detalles que nadie de son importantes instead of *los detalles que nadie ve son importantes, y los festival internacional* instead of *22 festival internacional, dividan enfermedad* instead of *debido a una enfermedad, vida el gobierno* instead of *pide al gobierno, es la queja historia* instead of *es la vieja historia*.
Italian:
di un vero e proprio piano Marshall sulla lì instead of *di un vero e proprio piano Marshall per la Libia, un'ala* instead of *un'aula, nella notte si legge che il Presidente ha proceduto* instead of *nella nota si legge che il Presidente ha proceduto.*
German:
sie haben Maschinenpistolen eisig getragen instead of *sie haben Maschinenpistolen bei sich getragen, geschossen werden immer kleiner* instead of *die Chancen werden immer kleiner, die Grünen haben ihren Ausdruck beschlossen* instead of *die Grünen haben ihren Austritt beschlossen.*

The following are examples of standard edition errors:

English:
Birmingham's problems aren't solely of the council's making. There is a large population living in some of the most deprived communities in the country instead of *Birmingham's problems aren't solely of the council's making. There is a huge demand for services in this city. There is a large population living in some of the most deprived communities in the country*. (Omission of an independent idea unit: "*There is a huge demand for services in this city*").
We'll be discussing tonight. *Then we'll have some time at the end for football* instead of *We'll be discussing tonight the great start of the new labour goverment*. *Then we'll have some time at the end for football*. (Omission of a dependent idea unit ("*the great start of the new labour goverment*") that renders the remaining unit ("*We'll be discussing tonight*") meaningless).
Spanish:
El Celta de Vigo ha cuajado una gran temporada. El Deportivo de la Coruña, sin embargo, sigue decepcionando instead of *El Celta de Vigo ha cuajado una gran temporada. Pocos medios lo han mencionado hasta ahora. El Deportivo de la Coruña, sin embargo, sigue decepcionando.*
El ministro anunció. *Nadie se lo esperaba* instead of *El ministro anunció que dejará la política a finales de año*. *Nadie se lo esperaba*
Italian:
L'iniziativa verrebbe respinta anche nella Svizzera italiana. Va però detto che nella Svizzera italiana le campagne iniziano sempre tardi instead of *L'iniziativa verrebbe respinta anche nella Svizzera italiana.*

[7] In this example ("She has no big plans for *hell of even* this year"), *hell of even* does not create a new meaning that could make sense in the context of the sentence, but rather makes it difficult for the viewers to understand what was originally meant.

Regione che però si è sempre mostrata molto critica nei confronti della libera circolazione. Va però detto che nella Svizzera italiana le campagne iniziano sempre tardi.
German:
Das neue Gesetz erlaubt den Anbau und Verkauf von Haschisch. Ebenso ist der Konsum von vierzig Gramm pro Monat gestattet instead of *Das neue Gesetz erlaubt den Anbau und Verkauf von Haschisch. Eine Behörde soll die Produktion und den Handel überwachen. Ebenso ist der Konsum von vierzig Gramm pro Monat gestattet.*

3.2.3. Minor errors

These errors allow viewers to follow the meaning/flow of the original text and sometimes even to reconstruct the original words. Typical cases of minor recognition errors are the presence/absence of capital letters, apostrophes, insertions of small words, etc. Minor edition errors often involve the omission of a dependent idea unit that does not render the remaining unit meaningless or nonsensical, i.e. the omission of "this morning" in "the blaze started this morning at the front of the house". Minor edition errors depend largely on specific respeaking practices. In some countries, such as in the UK, respeaking *the former head of the Federal Reserve, Alan Greenspan, has stated that...* as *the former head of the Federal Reserve has stated that* may be considered as an edition error, whereas for those adopting a less verbatim approach, as is the case in Switzerland, it may be regarded as a correct edition. Corrected errors may be included in this category too although in most countries they are considered as correct editions. From the viewers' point of view, minor errors may go unnoticed or may be detected without hindering the comprehension of the key elements of the original text.

Here are some examples of minor recognition errors in different languages:

English:
brown instead of *Brown*, *we're* instead of *were*, *it's a Ryan Giggs* instead of *it's Ryan Giggs*, *for people were found* instead of *four people were found*, *their* instead of *they're*, *what you do then?* instead of *what do you do then?*
Spanish:
va ser crucial instead of *va a ser crucial*, *ayer estudie mucho* instead of *ayer estudié mucho*, *el presidente zapatero* instead of *el presidente Zapatero*, *todo satisfechos* instead of *todos satisfechos*, *tan poco lo tiene* instead of *tampoco lo tiene*.
Italian:
credo instead of *crede*, *là* instead of *della*, *ne meno ha* instead of *nemmeno ha*, *se si sono cinque* instead of *se ci sono cinque*, *è caso* instead of *è il caso*, *il* instead of *gli*, *questa quello* instead of *questo è quello*, *napolitani* instead of *Napolitano.*
German:
zweite instead of *zweiten*, *diese* instead of *dieser*, *bereit haben* instead of *bereits haben*, *dass die irische Regierung zwei weitere konkursreifer Banken Verstaatlichung muss* instead of *dass die irische Regierung zwei weitere konkursreife Banken verstaatlichen muss.*

And the following are examples of minor edition errors:

English:
The neighbours did all they could to try and get inside, but it was too difficult and dangerous a task.
instead of
The neighbours did all they could to try and get inside, trying to knock down the door, but it was too difficult and dangerous a task.
Spanish:
La playa de las Islas Cíes fue elegida como la mejor del mundo en una encuesta publicada en 2007.
instead of
La playa de las Islas Cíes fue elegida como la mejor del mundo en una encuesta publicada por The Guardian en 2007.
Italian:
Sono emerse alcune novità nell'inchiesta sull'incidente di Michael Schumacher comunicate nel corso di una conferenza stampa a Grenoble
instead of

Sono emerse alcune novità nell'inchiesta sull'incidente di Michael Schumacher comunicate nel corso di una conferenza stampa a Grenoble dalla procura francese.
German:
Die Studie hat für Unruhe unter den Parlamentariern gesorgt.
instead of
Die Studie, die in der NZZ veröffentlicht wurde, hat für Unruhe unter den Parlamentariern gesorgt.

4. Application of the NER model in English

This section includes 3 examples of how the NER model can be applied to the assessment of accuracy in live subtitling, in this case with respoken subtitles.

Example 1

In example 1, although the respeaker has had to edit 20% of the original discourse due to the high speech rate, the key information has been included in the subtitles, which are very accurate:

Original text	Respoken subtitles
- What would it mean for triathlon in Britain if a male or female British athlete were able to go to Beijing and bring back a first triathlon Olympic medal?	- What would it mean for triathlon in Britain if a male or female British athlete were able to go to Beijing and bring back a first triathlon Olympic medal?
- Oh, I think it'd be, be, be everything for, for the individual, of course, and, you know, I'm sure, you know, there'd be loads of people, you know, all the personal supporters but obviously, you know, the sport as a whole would be great. You know, I'm sure that if you speak to the rest of the guys that's what everyone is really, really trying to do is get a good performance and, you know, try to race for a medal and, you know, it's the pinnacle of your career, you know and, you know, hopefully and if it's one of the younger ones, or myself or whoever, you know, it'll be a stepping stone to, to, to, for experience leading up to 2012.	- It would be everything for the individual, of course. I'm sure there will be lots of people who will... The sport as a whole would be great. I'm sure if you speak to the rest of the guys, that's what everyone is trying to do. Get a good performance and try to race for a medal. It is the pinnacle of your career. Hopefully if it is one of the younger ones or myself, whoever, it will be a stepping stone for experience leading up to 2012.
- Inevitably, we have to touch on the three missed tests and at the fact that you were temporarily banned by the BOA. Do you look back on it and feel that you were maybe a victim of a system that was in its infancy at the time?	- Inevitably, be have to touch on the three missed tests and at the fact that you were temporarily banned. Do you look back and it and feel that you were maybe a victim of a system that was in its infancy at the time?
- Uh, to a degree, yes. I think that everyone learnt a lot from that. UK sport did, I did, the federation and hopefully the juniors have. Uh, but, you know, it happened and, you know, I'm not hiding behind the fact it didn't happen.	- Everyone learnt a lot from that, UK sport, the federation and hopefully the juniors. It happened. I am not hiding behind the fact it didn't happen.

$$\text{Accuracy} \quad \frac{205 - 1 - 0.5}{205} \times 100 = 99.3\%$$

N: 205	(186 + 19 commands, namely commas, full stops and question marks)
E: 1	(*all the personal supporters* [0.25: dependent idea unit], *by the BOA* [0.25: dependent idea unit], *to a degree, yes* [0.5: independent idea unit]).
R: 0.5	(*be have* instead of *we have* [0.25], *look back and it* instead *look back on it* [0.25])
CE: 40	(*oh, I think* [x2], *be* [x2], *for, and* [x5], *you know* [x 12], *but obviously, that, really* [x2], *or, to* [x3], *I think that, I did, have, and, but, it would be, there will be, it is, it will be*)
Assessment:	Overall accuracy is very good. There is heavy editing (20.1% of the original text) due to very high speech rate (245 wpm) but, to the respeaker's credit, there are also 40 instances of correct editing vs. only 3 where some information is lost. The respeaker is very proficient in this regard. Recognition is good but attention should be paid to *be/we* and *and/on*. In any case, when respeaking with Dragon, if the respeaker manages to deliver respeaking units such as *we have to touch* or *look back on it*, the algorithm in the language model is unlikely to allow mistakes such as *be have to touch* and *look back and it*. None of the errors is serious.

Example 2

Example 2 presents a very different situation. In this case, only 6% of the original discourse has been edited, but while many irrelevant elements have been maintained in the subtitles, other key idea units have been lost.

Original text	Respoken subtitles
Everyone agrees the economy is gonna cool down in two years. The question is, will it be a deep freeze or just a bracing chill? Well, if I really knew the answer to that question, I'd be in the City, not standing here talking to you. But I do know what the key questions are: the big one is the US, and whether it's about to slip into recession.	Everyone agrees the economy is gonna cool down. The question is, will it be a deep freeze or just a bracing chill? Well, if I really knew the answer to that question, I'd be in the City, not standing here talking to you. But I do know what the key questions are, the big one is whether it's about to slip into recession.
The former head of the Federal Reserve, Alan Greenspan, thinks the odds on a recession this year are fifty-fifty. The Fed has cut rates three times this year but has been surprised by the slowdown and plans to do more. The same goes for President Bush, who said yesterday he was thinking about a stimulus package of his own for 2009.	The former head of the Federal Reserve thinks the odds on a recession this year are high. The Fed has cut rates but has been surprised by the slowdown and plans to do more. The same goes for President Bush, who said today he was thinking about a stimulus package of his own.
Now, the worse things get over there, the tougher it will be in Britain. Sure, people talk about all the growth in Asia and how the global economy can decouple from America, but Britain's credit crunch has been nearly as bad as America's. Worse, if you think Northern Rock. And even though Alistair Darling has announced reforms today that could prevent that kind of fiasco happening again at a macro level, it's hard to get round the fact that Britain shares many of the same big economic weaknesses as America, not least a habit of spending beyond our means.	Now, the worse things get over there, the tougher it will be in Britain. Sure, people talk about all the growth and how the global economy can decouple from America, but Britain's credit crunch has been nearly as bad as America's. Even though new reforms could prevent that kind of fiasco happening again, it's hard to get round the fact that Britain shares many of the same big economic weaknesses as America, not least a habit of spending beyond our mains.

$$\text{Accuracy} \frac{220 - 5.5 - 0.25}{220} \times 100 = 97.3\%$$

N: 220	(196 + 24 commands, namely commas, full stops and question marks)
E: 5.5	(*in two years* [0.25], *the US* [1], *Alan Greenspan* [0.25], *fifty-fifty* [1], *three times this year* [0.25], *today* [1], *for 2009* [0.25], *in Asia* [0.25], *Worse, if you think Northern Rock* [0.5], *Alistair Darling has announced* [0.25], *today* [0.25], *at a macro level* [0.25]).
R: 0.25	(*mains* instead of *means* [0.25])
CE: 1	(*and* [*even though*…])
Assessment:	Overall accuracy doesn't reach 98%. Recognition is good: only one minor error. The problem lies in editing, not because of the amount (only 6% edited, as opposed to over 20% in the previous example) but because of the quality (12 instances of incorrect editing vs. 1 instance of correct editing). Facing a fairly normal speech rate in the original text (165 wpm), the respeaker has kept many "irrelevant" elements (*well, but, sure, now, and*) but has lost as many as 12 idea units. Many of these are dependent idea units made up of only one or two words and could have easily been maintained. Further training is needed to improve this respeaker's editing skills.

Example 3

Finally, example 3 is also unsuccessful. In this case, though the problems are not related to edition, but to recognition, which means that the respeaker needs to perform further training with his/her voice profile.

Original text	Respoken subtitles
- In my view, it is monetary policy that needs to act, not fiscal policy. And what we have at the moment is interest rates which are contractionary on the economy. Let's not forget that an interest rate of 15% is higher than a neutral interest rate for this economy and yet there are all these forces pushing downwards, such as oil prices, the credit crunch, housing, housing vulnerabilities. In my view, the Bank of England needs to start cutting rates and start cutting them quickly. Now, if it doesn't, this economy is likely to stall next year.	- In my view, it is monetary police that needs to and not fiscal policy. What we have at the moment is interest rates which are confectionery on the economy. An interest rate of 50% is higher than a neutral interest rate for this economy and yet there are all these forces pushing downwards, such as the credit card, housing, housing and inabilities. The Bank of England needs to start cutting rates and start cutting them quickly. If it doesn't, this economy is likely to still the next year.
- So if you were still in the committee that is what you would be reckoning? - Basically, I will be voting for a cut next year.	- So if he were still in the committee that is what you would be reckoning? - I will be voting for a cup next year.
- She's absolutely right. Look, sure there is a risk of inflation, we're not in the Weimar Republic, we're living in Great Britain in the United States. There is pain to cutting interest rates. I mean, you risk a little inflation but the question is what would you rather risk, a little more inflation or a major slowdown? Given the risks, you best take a chance on inflation.	- She's absolutely right. Sure there is a risk of inflation, were not in the way my Republican, were living in Great Britain in the United States. There is pain to cutting interest rates. You risk a little inflation but the question is what would you rather risk, a little more inflation or a major slowdown? Given the risks, you best take a chance on inflation.
- I was just talking to some people at the White House today. It looks like we have room for fiscal policy adjustments. The president is considering increasing allowances for depreciation for instance to stimulate business investment. In this country, Gordon Brown did not leave you with that room.	- I was just talking to some people at the White House today. It looks like we have room for fiscal policy adjustments. The president is considering increasing allowances for deposition for instance to stimulate business investment. In this country, Gordon Brown did not leave you with that run - room.

$$\text{Accuracy} \frac{259 - 0.25 - 8.75}{259} \times 100 = 96.5\%$$

N: 259 (232 + 27 commands, namely commas, full stops, question marks and one correction)

E: 0.25	(*oil prices* [0.25])
R: 8.75	(*police* instead of *policy* [1], *and* instead of *act* [0.5], *confectionary* instead of *contractionary* [0.5], *50%* instead of *15%* [1], *credit card* [1], *and inabilities* instead of *vulnerabilities* [0.5], *to still* instead of *to stall* [0.5], *he* instead of *you* [1], *cup* [0.5], *were* instead of *we're* [0.25], *were* instead of *we're* [0.25], *way my Republican* instead if *Weimar Republic* [0.5] and *depreciation* as *deposition* [1], *run* corrected as *room* [0.25])
CE: 7	(*and, let's not forget that, in my view, now, basically, look, I mean*)
Assessment:	Overall accuracy doesn't reach 98%. Editing is good: only 1 dependent idea unit missing and 7 instances of correct edition. Recognition is poor: 14 recognition errors (including 5 serious errors). Further training is needed to improve the voice profile. The errors occur not only with single words but also phrases and contractions. The respeaker should thus be advised not to dictate contractions in order to avoid errors such as *were* instead of *we're*.

5. Final thoughts: the NER model and automatic subtitling

As mentioned above, the NER model can also be used for automatic subtitles, that is, those produced with ASR (automatic speech recognition). Given the rapid evolution of speaker-independent speech recognition, it makes sense to anticipate that subtitling companies will try to resort to this technology once it has reached optimum levels of accuracy. First of all, this software may be used with the intervention of a human operator, who will correct misrecognitions and errors of punctuation and speaker identification before sending the subtitles on air. And perhaps in a more distant future, human intervention may be excluded from the process altogether. Be that as it may, it is important to ensure that these subtitles are at least as accurate as those produced by respeaking, in this case reaching a 98% with the NER model, including punctuation and character identification.

Yet, there is one more element that becomes critical when dealing with automatic subtitles: speed. As highlighted in Romero-Fresco (2012), the speed of the subtitles has a direct impact on the amount of time viewers can devote to the images. According to the eye-tracking data obtained in Poland, the UK and Spain in the DTV4ALL project, and in South Africa by Hefer (2011), a speed of 150wpm leads to an average distribution of 50% of the time on the subtitles and 50% on the images. A faster speed of 180wpm yields an average of 60-65% of the time on the subtitles and 40%-35% on the images, whereas 200wpm only allows 20% of the time on the images. As shown in González Lago (2011), the average speech rate of live programmes such as the Spanish news is 240-278 wpm, with peaks of 400wpm and even 600 wpm in certain cases. Speech rates of over 220wpm are also very common for presenters in the UK (Eugeni 2009). Considering that these presenters are unlikely to slow down their speech rates and that automatic subtitles are by definition verbatim, the speed of automatic subtitles is likely to cause viewers to miss most of the images, unless a) the human intervention before launching the subtitles also includes edition, which is very complex and could lead to great delays, b) an antenna delay is implemented so that the editor can have time to correct errors and edit the subtitles[8] or c) the technology used allows to define settings in order to achieve an optimum display mode/exposure time of subtitles by automatically calculating a maximum and minimum duration.

6. Conclusions

Following a brief description of some of the traditional models used to assess accuracy in speech recognition and respeaking, the present article has introduced the NER model. This is an attempt to provide a functional and easy-to-apply model that can assess the accuracy of live subtitles, while also providing data on important subtitling factors such as delay, position, speed and character identification. The division between edition and recognition errors allows the trainer to have not only a figure indicating the accuracy rate of the subtitles, but also an idea of what must be improved and how it can be improved.

At the time of writing, the model has been adopted by Ofcom to assess the quality of live subtitles on UK TV (Ofcom 2013) and has been included in the official Spanish guidelines on subtitling for the deaf and hard of hearing (AENOR 2012). It is also being used by broadcasters, companies and training institutions in Spain, France, Italy, Switzerland, Germany, Belgium and Australia, among other countries. Furthermore, NERstar, a semi-automatic tool, has been developed to ensure a quick and effective application of the NER model to live subtitles produced by respeaking or by automatic speech recognition (ASR).

10 years after the introduction of respeaking, the collaboration between academia and the industry seems to be yielding interesting results, not least a set of parameters to ensure a certain degree of quality in

[8] This is done in Belgium by the Flemish public broadcaster VRT, which has managed to implement an antenna delay of up to 10 minutes for some live programmes such as chat shows. This allows subtitlers and respeakers to correct, edit and synchronise the subtitles before launching them on air for the viewers, who receive them as though they were produced live.

respeaking, which in this case could include 98% accuracy with the NER model, a block display mode and a maximum speed of 180wpm. Now that the use of (semi)automatic subtitles is becoming a possibility, it is all the more important to maintain these parameters in order to ensure that these new developments are not introduced at the expense of the quality of live subtitling, that is to say, at the expense of the viewers.

7. References

AENOR (2012) *Subtitulado para personas sordas y personas con discapacidad auditiva.* Madrid: AENOR.

Apone, Tom, Marcia Brooks and Trisha O'Connell (2010) *Caption Accuracy Metrics Project. Caption Viewer Survey: Error Ranking of Real-time Captions in Live Television News Programs*, report published by the WGBH National Center for Accessible Media. http://ncam.wgbh.org/invent_build/analog/caption-accuracy-metrics

Chafe, Wallace (1985) "Linguistic differences produced by differences between speaking and writing", in David Olson, Nancy Torrance and Angela Hildyard (eds.) *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing*. Cambridge: Cambridge University Press, 105-122.

Dumouchel, Pierre, Gilles Boulianne and Julie Brousseau (2011) "Measures for quality of closed captioning", in Adriana Şerban, Anna Matamala and Jean-Marc Lavaur (eds) *Audiovisual Translation in Close-up: Practical and Theoretical Approaches*, Bern: Peter Lang, 161-172.

Eugeni, Carlo (2009) "Respeaking the BBC News. A strategic analysis of respeaking on the BBC". *The Sign Language Translator and Interpreter* 3(1): 29-68.

González Lago, María Dolores (2011) *Accuracy Analysis of Respoken Subtitles Broadcast by RTVE, the Spanish Public Television Channel*. MA Dissertation. London: Roehampton University.

Hefer, Esté (2011) *Reading Second Language Subtitles: A Case Study of Afrikaans Viewers Reading in Afrikaans and English*. MA Dissertation. Vaal Triangle Campus of the North-West University.

Ofcom (2013) *Measuring the quality of live subtitling: statement*, London: Ofcom.

Romero-Fresco, Pablo (2011) *Subtitling through Speech Recognition*: *Respeaking*, Manchester: St Jerome.

— (2012) "Quality in live subtitling: the reception of respoken subtitles in the UK", in Aline Remael, Pilar Orero and Mary Carroll (eds.) *Audiovisual Translation and Media Accessibility at the Crossroads.* Amsterdam: Rodopi, 111-131.