Respeaking in minority languages: Development of a Catalan automatic speech recognition system

Estel·la Oncins, TransMedia Catalonia[1], Universitat Autònoma de Barcelona, Spain
Héctor Delgado, Department of Digital Security, EURECOM, France

This contribution aims to describe the current situation of accessibility services in live events, taking stock of the latest tendencies in the elaboration of real time intralingual subtitles in live events at the Universitat Autònoma de Barcelona (UAB). Secondly, a description of the development of a system architecture, initial set up of the automatic speech recognition (ASR) system in Catalan and the results obtained will be provided. Finally, we will outline the advantages that Internet-based technologies could provide to improve accessibility services to audiences. Last section will conclude the paper and raise questions for future research.

Keywords: *Audiovisual translation, Accessibility, real time intralingual subtitling, live events, respeaking, automatic speech recognition (ASR), minority language, speech to text technologies*

1. Introduction

Rendering live events, such as conferences and meetings accessible in real-time to deaf and hearing-impaired audiences is becoming increasingly possible in all countries, primarily thanks to the advances made in speech recognition technologies. But in the case of minority languages such as Catalan, no commercial speech recognition programme is available on the market. Therefore, stenography-based technologies are still used. However, nowadays the lack of professional stenographers is becoming an increasing problem. This restricted accessibility of information to Catalan deaf and hearing-impaired audiences is in part due to a lack of available technologies in minority languages.

Given the many and varied types of live events, this work does not aim to provide an ontological conceptualisation of what a live event is, but rather a contingent definition of how it can be rendered accessible to all audiences. Despite the intrinsic differences all audiovisual events share one principle: they have an audience that needs synchronized access to both, verbal and visual information. Therefore, focus will be placed in live events that take place at the university and in particular: course inaugurations, academic conferences and homages. When developing the project proposal the following three main elements have been taken into account: audiences, typology of the event and available facilities of the venue.

1.1. Audiences

If we take a look at the national and international scene as Chaume (2013:107) highlights 'policies of equality and media accessibility, have spawned a row of new

[1] "TransMedia Catalonia is a research group funded by Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya, under the SGR funding scheme (ref. code 2017SGR113)"

audiovisual translation modes, designed to meet the variety of needs or concerns of different social groups'. In fact, according to WHO [2] over 5% of the world's population – or 466 million people – has disabling hearing loss. In 2050 this figure is expected to increase to over 900 million people, which means that one in every ten people will have disabling hearing loss. Against this background, several laws have been approved at the autonomic, national and European level to ensure that this group has equal accessibility without obstacles to the information that is being provided in live events. Within this context, the adoption of the United Nations Convention on the Rights of Persons with Disabilities[3], claims access to information as a basic human right. In addition, the agreement on the European Accessibility Act[4] and the Audiovisual Media Service Directive (AVMSD)[5], governments throughout Europe are being forced to make major efforts to increase the accessibility services of their media, administration and institutions. The most obvious and regulated case with regard to live subtitling is that of the media, specifically television.

It is difficult to establish exact taxonomies and classifications for the study of subtitling live events outside the TV, such as sessions of the Parliament, awards ceremonies, homages, conferences, interviews, discussions, and many others, which in some cases can take place indoors and outdoors. In the following paper the focus is placed on live events that take place at the university and in particular: course inaugurations, academic conferences and homages. Despite all their differences, all these live events present one thing in common: they have an audience that needs access to the information.

The core question is: What is the point of rendering a live event fully accessible online, when sensory impaired audiences sitting in the auditorium have limited access to the media content of the live event in their own language? The social and technological advances of recent years are being crucial in making live events accessible for the deaf and hearing-impaired audiences.

1.2. Typology of the event

In terms of the discursive structure live events primarily differ from stage performances, such as: theatre and opera, because the latter presents a close structure in the form of a screenplay, which has been rehearsed and will frequently be maintained (Oncins 2013). Additionally, live events are subjected to further critical points, which are out of the scope of this paper, but are close related to the public speaking skills of the speaker, such as fast delivery pace of speakers, the use of regional dialects or the facilities of the venue, such as wrong audio set up. However, most live events held at the university present a common speech structure: greeting protocol (welcome address of the organizers and the authorities), presentation of the speaker followed by a personal speech composed of a ready-made part and an improvisation part.

[2] http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (last accessed 11/09/18)

[3] http://www.un.org/disabilities/documents/convention/convention_accessible_pdf.pdf (last accessed 21/11/18)

[4] https://ec.europa.eu/social/main.jsp?catId=1202 (last accessed 22/11/18)

[5] https://ec.europa.eu/digital-single-market/en/audiovisual-media-services-directive-avmsd (last accessed 22/11/18)

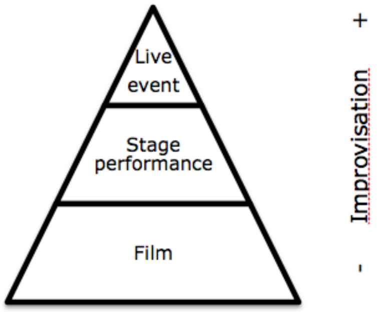

Figure 1. Improvisation comparison between different AV products

According to these AV products, Orero (2006) differentiates between three methods of creating subtitles: prepared subtitling (pre recorded AV products), semi-live subtitling (stage performances, TV news, etc.), and real time subtitling (live events). It should be mentioned that real time subtitling is usually done intralinguistically and aimed at deaf and hearing impaired audiences, but can also be used by linguistically impaired audiences. In terms of process, real time subtitling follows a different workflow than prepared and semi-live subtitling due to the timing constraints and load of speech. Therefore, reformulation and edition strategies are used.

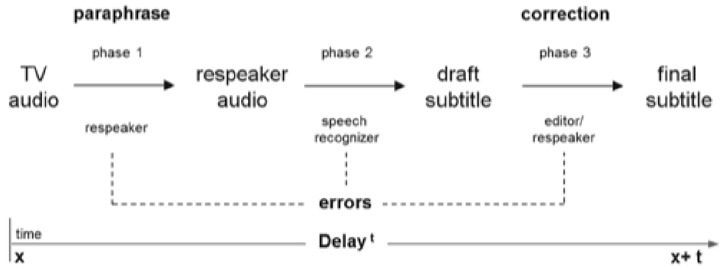

Figure 2. Temporal representation of the live block subtitling process based on Van Waes, Leijten & Remael 2013)

As it can be observed, synchronicity is a key issue, and perhaps the one, which poses the greatest challenge because it has a direct implication in the reception and perception of the final subtitle delivered to the audience. Within this context, the available facilities of the venue will be crucial.

1.3. Facilities of the venue

The Aula Magna of the Rectorate building at the Universitat Autònoma de Barcelona is equipped with a subtitling and audiodescriber workstations. Since the aim of this paper is real time intralingual subtitling, focus will be placed in the subtitling workstation, which in terms of hardware has a soundproof booth equipped with a computer and a headset microphone.

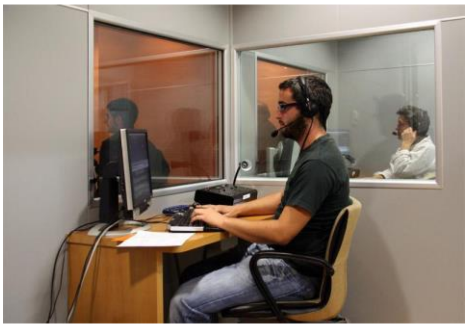

Figure 3. Subtitling workstation at the Rectorate building of the UAB

The computer has a respeaking platform installed for live subtitling that includes a ASR system in Catalan, but if a conference combines two or three different languages a speaker-dependent speech recognition software trained in Spanish or English can also be added.

The technician workstation is equipped with a TriCaster mixer, which allows broadcasting the event in a single image, combining the speaker, the sign language interpreter and the real time intralingual subtitles.

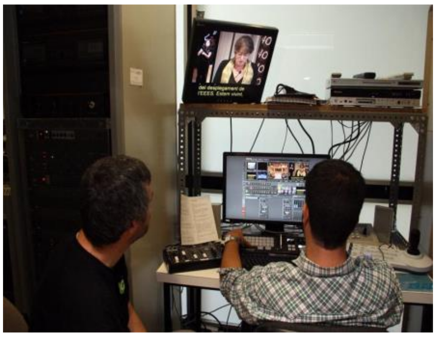

Figure 4. Technician workstation

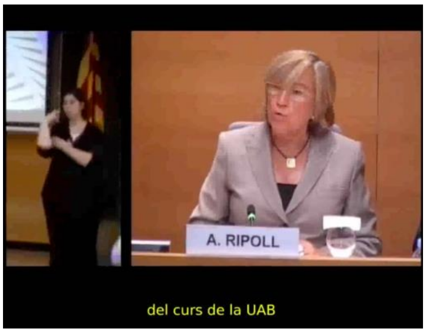

Figure 5. Tricaster mixer

Finally the hall is equipped with two video cameras and an open info accessibility screen connected to a projector. Since 2011, the live events have also been streamed online and a wireless system was implemented to deliver the available accessibility services to mobile devices[6].

2. Traditional and new real time intralingual subtitling techniques

Even nowadays, most live events and specially those held in minority languages, are made accessible to deaf and hearing-impaired audiences through professional stenographers. There is an existing lack of professionals in this field and training time is long and costly compared to respeaking or other techniques (see table 1).

[6] Further information regarding the system is provided in the article Oncins, E. *et al.* (2013). “Multi language and multi system mobile application to make accessible live performing arts: All Together Now”.

	Velotype	Tandem	Stenograph	Respeaking
Delay	Medium	Medium	Low	Low
Speed	120-150 wpm (when combined)	140-180 wpm (when combined)	up to 220 wpm	140-160 wpm
Accuracy	95-98%	95-98%	97-98%	95-98%
Difficulty	Low	High	Very high	Medium
Cost	Medium	Medium	High	Low
Problems	High	Low	Medium	Medium

Table 1. Comparison of the main real time intralingual subtitling systems[7]

As it can be observed in table 1, in terms of cost and training, respeaking is the less expensive subtitling technique and the most preferred one nowadays (Romero Fresco 2018). However, since it is still a speaker-dependent technique it relies on the number of languages available and minority languages like Catalan are yet not covered.

2.1. Catalan ASR system

The central reason that lead TransMedia Catalonia research group to collaborate in the development of an ASR system in Catalan was to cover the lack of a speech recognition engine in this language.

The following subsections give details of this ASR system used by the respeaking platform. Any standard ASR system consists of two main components. First, the acoustic modelling that learns the different sounds of the language (commonly at phone level). A big amount of speech data with its word-level transcriptions is required in order to map sounds with their phoneme-level representation. Once the phone-level models are trained, word-level models can be obtained with the concatenation of their phone unit models. This word model generation is possible thanks to a dictionary, which contains the phone transcription of all the words in the vocabulary. Second, the language modelling learns about the typical word sequences in the language being modelled. This module leads the ASR system to explore only the most likely word sequences instead of performing a brute force search among all possible word sequences. The language model is learned from big collections of text in the target language. Both acoustic and language models work together to jointly map an unknown speech input into the transcription hypothesis.

The following subsections will offer a more technical description in order to explain how the ASR components are obtained. First, the training data will be described. Then the acoustic and language models training will be detailed. Finally, the performance of the resulting ASR system will be assessed.

2.1.1 Training and test data

[7] Based on Lambourne (2006)

The Speecon Catalan speech corpus (Moreno, A. *et al.*, 2006) was employed in order to train the audio-based components of the system, namely the acoustic modelling. This database contains recordings from 550 adult speakers balanced in gender and in Catalan dialect, including central, "Nord-occidental", "Gironí" and "Tortosí". Each utterance was recorded with up to four microphones simultaneously located at different distances. The speech recordings were sampled at 16 kHz. Every session consists of 291 read utterances plus 30 spontaneous spoken utterances. Examples of the content of such utterances are free 5-minute spontaneous speech, short spontaneous utterances (i.e. dates, hours, proper nouns, cities, telephone numbers, etc), basic read words and sentences, application words, and phonetically rich sentences and words. The different sessions were recorded in a variety of environments such as offices, homes, cars and public places. The database contains orthographic annotations, a list containing the full lexicon and the corresponding phonetic transcription.

Transcriptions of the plenary sessions of the Catalan Parliament, consisting of around 24 million words were used for language modelling. The Parliament minutes were downloaded from the official website[8] in PDF format, which were automatically parsed to derive plain text, clean sentences. The complete 167.000 word vocabulary was reduced to the most common 64.000 words.

To test the recogniser, an excerpt from a Parliament session of 13 minutes was used. The original 48 kHz signal was down sampled to 16 kHz. Note that the test set differs in nature from the training set.

To train and test the ASR system, a traditional mel frequency cepstral coefficients (MFCC) frontend was employed. 12 static coefficients plus the 0-*th* coefficient were augmented with their first and second time derivatives, resulting in 39-dimensional feature vectors, which were extracted using a 25ms window with a 10ms shift.

2.1.2. Acoustic modelling

The acoustic modelling is based upon hidden Markov models (HMM) with Gaussian mixture density functions. First, a set of monophone models was obtained. We refer to monophones as the minimum sound units of a language. In other words, monophones are physical, acoustic examples of phonemes. A set of 39 monophone plus 1 silence HMMs, were flat initialised. Those implement the traditional, left-to-right 3-state topology with self-loops. Then the initial models were re-estimated on the speech training data with its phone-level transcriptions through a few iterations of the Baum-Welch (BW) algorithm. This algorithm estimates the HMM parameters, namely the state priors, the transition probabilities, and the Gaussian continuous density functions. Later, a new "short pause" (SP) model was derived from the silence HMM by cloning its central state and by adding a "skippable" transition. The SP model was to be inserted between word boundaries to model possible short, between-word pauses.

Next, the monophone HMMs were used to leverage a set of triphone HMMs, where triphones are sequences of 3 monophones. This is usually done to achieve a smoother modelling of transitions of phones within the words. The phone-level transcriptions were processed to derive crossword triphone transcriptions by inserting the "sp" label, which refers to a potential short pause, and by grouping the monophones in sets of 3. HMMs for all triphones contained in the training data were

[8] https://www.parlament.cat/

then synthesised using the monophone HMMs. Every triphone HMM with a common central phone shares the central state parameters. A few iterations of the BW algorithm were performed again to re-estimate the model parameters on the training data.

Since the obtained triphone set does not cover all possible triphones in the language, the missing ones were synthesised as "logical models" which are "tied" to some already existing "physical model". This contributes to the overall robustness of the acoustic models. The tying process relies on a decision tree-clustering algorithm, which uses linguistically motivated questions about triphone contexts. The resulting tied-state triphones were re-estimated again through several BW iterations.

The emission probability functions were then improved by deriving multivariate Gaussian Mixture models from the original multivariate single-Gaussian models. A splitting approach was adopted to increase the number of components by powers of 2, followed by intermediate BW iterations, until reaching 32 Gaussian components.

Finally, the set of models was discriminatively trained to maximise the mutual information (e.g. maximum mutual information criterion, MMI). This was accomplished through an extended Baum-Welch (EBW) algorithm.

2.1.3. Language modelling

The language model is a 64k word, 3-gram model learned on the text training data described above. The Turing-Good discounting approach with cut-off (to discard infrequent events of the training data) was used to obtain a standard 3-gram model.

2.1.4. Experimental results

Two experiments were performed to benchmark the speech recogniser. Although experiments were conducted on a limited amount of test data, the results serve as an indicator of the performance under two different conditions, namely clean and noisy. The clean condition includes speech that has been recorded in a silent office. This condition matches somehow the conditions of a re-speaking application in which a silent, acoustically treated room is expected to be available. The noisy condition, however, corresponds to a much more challenging scenario in which a direct transcription is performed from the speaker during the live event. This latter condition includes a speech recording captured at the Catalan Parliament, extracted from the videos publicly available at the Parliament website. The audio files were processed with the ASR system to obtain the 1-best word-level hypothesis. The hypotheses were then compared with the reference word-level transcriptions in order to compute the word error rate (WER). This is by far the most extended metric to assess speech recognition performance, and calculates the percentage of words on the hypothesis that these are incorrect with regard to the ground-truth reference. Two different model configurations were evaluated: maximum likelihood (ML) training, and discriminative training with maximum mutual information criterion (DT-MMI). Also, the impact of speaker adaptation through maximum likelihood linear regression (MLLR) was assessed. Table 2 shows the experimental results.

	Clean speech	Noisy speech

	No speaker adaptation	With speaker adaptation	No speaker adaptation	With speaker adaptation
ML	20.95	17.72	28.99	25.01
DT-MMI	20.21	16.14	27.95	24.90

Table 2: Automatic speech recognition performance in terms of word error rate (WER, %), for two different training approaches (maximum likelihood, ML, and discriminative training with maximum mutual information criterion, DT-MMI), and with and without speaker adaptation. Results are provided for clean (office recordings) and noisy (Parliament recordings) conditions.

Results show that discriminatively trained acoustic models (DT-MMI row) universally outperform the acoustic models trained with maximum likelihood (ML row). Furthermore, and somehow expected, performance under the clean condition is better than under the noisy condition. Finally, speaker adaptation leads to the best performance of 16.14% and 24.90% WER on clean and noisy conditions, respectively. Taking into account that the clean condition with speaker adaptation is perfectly reasonable for a respeaking application (i.e. with controlled acoustic conditions and with a fixed respeaker), and also taking into account the limited resources in Catalan language employed for system development, the system performance with a 16.14% WER is encouraging. It would be expected that performance would further improve by increasing the amount of training data. Furthermore, the use of more recent, state-of-the-art acoustic modelling technology based on deep neural networks would bring even better accuracy.

Finally, note that this assessment considered the *offline* processing of the test speech files. To use the ASR system for real-time applications (i.e. respeaking), the offline decoder is replaced with an *online* decoder that processes the audio stream captured from an input microphone in real time.

2.2. Catalan ASR in the subtitling workstation

The ASR system was implemented in the subtitler workstation in 2011. In terms of workflow, the respeaker listens to the speaker source text (ST) and respeaks to a microphone connected to the Catalan ASR, which processes and delivers the spoken target text (TT) in a subtitler editor [9]. The text can be modified/corrected before delivering the end subtitles to the open screen and to the Wi-Fi for smartphone delivery.

[9] The subtitler editor was developed following the UNE 153010:2012 guidelines (Subtitling for deaf and hard-of-hearing people)

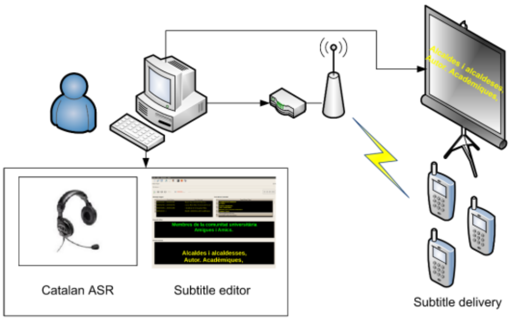

Figure 6. Subtitling system at the Rectorate building

In terms of performance of the system the following improvements compared to the stenography were found:

- Better readability for the audience
- Higher efficiency rates for the subtitler
- Elimination of the typing mistakes
- Available transcription of the subtitles to be used for the off line version of the video

Despite the progresses showed, the following adverse issues must be taken into account:

- 6-10 seconds delay
- 90% accuracy rate in ideal conditions

To increase the quality and accuracy rates of the ASR system, it is crucial to further develop the acoustic and language modelling in order to reduce the delay and optimize synchronicity.

3. Internet-based technologies and new platforms

It could be asserted that the future of real time intralingual subtitling is closely linked to respeaking. The rapid development of speech recognition technology in the following years will be crucial for the future of professionals in this field. Still, some fundamental issues to cover from the technology perspective are the need for an extreme audio controlled environment, which most of live events outside the TV, cannot offer and the voice modulation of the speaker.

3.1. Automatic speech recognition technologies

As technology evolves, the speech recognition engines, and specifically the speaker-independent systems will continue to improve their accurate rates, which remains a major challenge (see table 3). Some systems have already reported very good results under controlled conditions.

Characteristics	Speaker Dependent Recognition	Speaker Independent Recognition
	Requires time consuming user training. Flexibility in	Requires no a priori user training.

Convenience	changing users is reduced.	
Accuracy	Accuracy is higher due to available information on user's voice.	Accuracy suffers from lack of specific data and depends on the quality of the audio.
Robustness	Performance deteriorates as user's voice changes from training tokens.	Speaker independent recognition is robust to variations in speech.
Availability	Low cost speaker dependent systems are available today.	Few speaker independent systems are available and open source.

Table 3: Speaker dependent vs. speaker independent recognition, based on Karlsson 1990

Rendering real-time live events accessible to audiences is becoming increasingly possible in all countries even in minority languages, such as Catalan. One clear example is webcaptioner (www.webcaptioner.com), a speaker-independent ASR open source project that offers captioning services in over 40 languages, including Catalan. However, the quality of the output text produced by the speech recognition engine is highly dependent not only on the modulation of the voice, speech pace, available data and discourse cohesion of the speaker but also on the background noise. Additionally, regional dialects, punctuation marks, personal or company names cannot be recognized.

4. Conclusions

Speech-to-text technologies have been researched for a long time now and technological developments especially in speaker-dependent languages are already showing high quality results. Still in the case of minority languages, like Catalan, the developments are relatively recent and the accurate rate very poor. It should be highlighted that these technologies were not initially conceived for media accessibility purposes, therefore closer cooperation between audiovisual translators and engineers is crucial in order to align technology with real time subtitling needs, increase quality rates, and optimize synchronicity.

According to Romero-Fresco (2018), in a possible near future real time subtitlers may become editors of automatically recognized subtitles that they correct and cue live or may disappear altogether if broadcasters decide to show live subtitles produced by automatic speech recognition without any editing or human intervention. Therefore, research on quality and close cooperation with engineers will thus be essential to ensure that these automatic subtitles meet the standards required by the viewers.

This paper presents an example of such cooperation between media access experts, and engineers towards the development and testing of a new experimental ASR in Catalan language. Preliminary results show a low performance of the speech recognition systems, which may impede a professional application of this application at this stage, but it clearly leaves room for further development intended to cater for accessibility needs for both linguistically and sensory impaired audiences when attending a live event.

References

AENOR. 2012. *Subtitulado para personas sordas y personas con discapacidad auditiva. Subtitulado a través del teletexto, Norma UNE 153010:2012.* Madrid: Asociación Española de Normalización y Certificación (AENOR), 2012.

ARMA, Saveria. 2015. Real time subtitling for the Deaf and Hard of Hearing: An introduction to conference respeaking. In *Subtitling today: shapes and their meanings* (Perego, E., & Bruti, S., Eds.). Newcastle upon Tyne, UK: Cambridge Scholars Publishing, 2015, pp. 119-134

ARUMÍ RIBAS, Marta, ROMERO-FRESCO, Pablo. 2008. A practical proposal for the training of respeakers. *Journal of Specialised Translation, 10* [online]. 2008, pp. 106–127. Available at: <http://www.jostrans.org/issue10/art_arumi.php>

BARTOLL, Eduard. 2012. *La subtitulació: Aspectes teòrics i pràctics*. Vic: Eumo Editorial, 2012.

CHAUME, Frederic. 2013. The Turn of Audiovisual Translation. New Audiences and New Technologies. In *Translation Spaces 2*, 2013, pp. 105–123.

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

KARLSSON, Joakim. 1990. The integration of automatic speech recognition into the air traffic control system. *Massachusetts Institute of Technology*.

LAMBOURNE, Andrew. 2006. Subtitle respeaking. A new skill for a new age. In *inTRAlinea: online Translation journal.* [online].
Available at: <http://www.intralinea.org/specials/article/Subtitle_respeaking>

MORENO, Asunción, FEBRER, Albert, MÁRQUEZ, Lluís. 2006. Generation of Language Resources for the Development of Speech Technologies in Catalan. *Proc. of the Language Resources and Evaluation Conference LREC 06.* Genova: Italy. 2006.

ONCINS, Estella, LOPES, Oscar, ORERO, Pilar, SERRANO, Javier. (2013). Multi language and multi system mobile application to make accessible live performing arts: All Together Now. In *JosTrans (Issue 20): Image, Music, Text…? Translating Multimodalities.* [online]. Available at: <http://www.jostrans.org/issue20/art_oncins.php>

ONCINS, Estella. 2013. The tyranny of the tool: surtitling live performances. In *Perspectives 23 (1),* 2013, pp. 42-61.

ORERO, Pilar. 2006. Real Time Subtitling: A Spanish overview. In *inTRAlinea: online Translation journal.* Available at:
<http://www.intralinea.org/specials/article/Real-time_subtitling_in_Spain>

ROMERO-FRESCO, Pablo. 2018. Respeaking: Subtitling through Speech Recognition". In *The Routledge handbook of audiovisual translation studies*. (Pérez-González, L., Ed.). London: Routledge, 2018.

ROMERO-FRESCO, Pablo. 2009. More haste less speed: Edited vs. verbatim respeaking. In *Vigo International Journal of Applied Linguistics, VI*, 2009, pp. 109–133.

VAN WEAS, L., Leijten, M., and Remael, A. 2013. Live subtitling with speech recognition: causes and consequences of text reduction. In *Across Languages and Cultures, 14 (1)*, 2012, pp. 15-46.

In SKASE Journal of Translation and Interpretation [online]. 2019, vol. 12, no. 2 [cit. 2019- 20-12]. Available online at
http://www.skase.sk/Volumes/JTI17/pdf_doc/03.pdf. ISSN 1336-7811.