TRANS · núm. 24 · 2020
MISCELÁNEA · 263-281

Although interpreting has not yet benefited from technology as much as its sister field, translation, interest in developing tailor-made solutions for interpreters has risen sharply in recent years. In particular, Automatic Speech Recognition (ASR) is being used as a central component of Computer-Assisted Interpreting (CAI) tools, either bundled or standalone. This study pursues three main aims: (i) to establish the most suitable ASR application for building ad hoc corpora by comparing several ASR tools and assessing their performance; (ii) to use ASR in order to extract terminology from the transcriptions obtained from video-recorded speeches, in this case talks on climate change and adaptation; and (iii) to promote the adoption of ASR as a new documentation tool among interpreters. To the best of our knowledge, this is one of the first studies to explore the possibility of Speech-to-Text (S2T) technology for meeting the preparatory needs of interpreters as regards terminology and background/domain knowledge.

KEY WORDS: Speech-to-Text, computer-aided interpreting tools, terminology extraction, automatic speech recognition, ad hoc corpus, interpreting technologies.

Speech-to-Text Technology as a Documentation Tool for Interpreters: a new approach to compiling an *ad hoc* corpus and extracting terminology from video-recorded speeches

Mahmoud Gaber
Universidad de Málaga
Gloria Corpas Pastor
Universidad de Málaga
Ahmed Omer
University of Wolverhampton

La tecnología habla-texto como herramienta de documentación para intérpretes: Nuevo método para compilar un corpus ad hoc *y extraer terminología a partir de discursos orales en vídeo*

Aunque el ámbito de la interpretación no se ha beneficiado de los desarrollos tecnológicos en la misma medida que en traducción, actualmente asistimos al surgimiento de gran interés por desarrollar soluciones adaptadas a las necesidades de los intérpretes. En concreto, el Reconocimiento Automático de Habla (RAH) comienza a ser utilizado como parte de las herramientas de interpretación asistida, bien como componente de tales sistemas o como aplicación autónoma. El presente estudio persigue tres objetivos principales: i) determinar la herramienta de transcripción automática más apropiada para la compilación de corpus ad hoc*, comparando diversos sistemas de transcripción automática y evaluando su rendimiento; ii) utilizar RAH para extraer terminología a partir de las transcripciones de discursos orales en vídeo; y iii) promover el uso de RAH como nueva herramienta documental en interpretación. Se trata de uno de los primeros estudios en los que se abordan las posibilidades que ofrece la tecnología habla-texto para cubrir las necesidades terminológicas y documentales de los intérpretes en la fase de preparación de un encargo dado.*

PALABRAS CLAVE: *transcripción automática, herramientas de interpretación asistida por ordenador, extracción de terminología, corpus* ad hoc*, tecnologías de la interpretación.*

RECIBIDO EN FEBRERO DE 2020 ACEPTADO EN JUNIO DE 2020

1. INTRODUCTION

Nowadays, there is widespread agreement that high-quality interpreting is contingent upon advance preparation of the assignment. According to Fantinuoli (2017a: 24): "Preparation has been proposed in the literature as one of the most important phases of an interpreting assignment, especially if the subject is highly specialised". Interpreting requires efficient use of highly domain-specific terminology (in all working languages involved) with very limited time to prepare new topics. Besides, interpreters are frequently faced with variation in terminology, especially due to the fact that they are dealing with spoken language. For these reasons, corpus-based terminological preparation is rapidly gaining ground among professional interpreters and interpreter trainers (Pérez Pérez, 2017; Xu, 2018; Fantinuoli and Prandi, 2018; Arce Romeral and Seghiri, 2018, etc.).

However, gathering corpora for interpreting is not always an easy task. Leaving aside some negative attitudes towards technology and/or low degrees of technology uptake, interpreters usually experience an acute shortage of relevant resources and written materials. This is the situation with under-resourced languages (e.g. Urdu, Wolof, Khazakh), new developing domains, latest discoveries and inventions, and for certain communication settings (medical interventions, refugees and asylum seekers interviews, business meetings, education and training course sessions, etc.). In such cases, acquiring terminology and subject knowledge prior to interpreting usually requires the transcription of spoken speeches (video or audio files). And, since spoken language differs from written language, professional interpreters are always keen to listen to spoken speeches during the documentation phase to familiarise themselves with the speaker's accent, common expressions, specific formulae, etc. The aforementioned reasons make spoken speeches an indispensable and valuable source of knowledge and documentation.

Speech-to-Text (S2T), also known as Automatic Speech Recognition (ASR) or computer speech recognition, refers to the process of converting the speech signal into a sequence of words using algorithms implemented in a computer (Deng and O'Shaughnessy, 2003). There are two main types of ASR engines: speaker-dependent and speaker-independent systems. The first one operates by learning the characteristics of a specific person's voice. In this case, training is always needed to enable the software to understand any new user's voice before starting the speech recognition. The second one recognises anyone's voice as it is already trained by using a large amount of data from many speakers.

Considering the potential of ASR systems for those communication cases where the written source is unavailable or insufficient, this research aims to take advantage of such technology and contribute to enhancing its integration into Computer-Aided Interpreting (CAI) tools. Transcribing speeches automatically may supply interpreters with precious "raw material" for corpus building, which can be exploited in various ways.

Unlike translators, interpreters seem to be somewhat reluctant to embrace technology in their daily work. In contrast with heavily technologised translation environments, interpreters have a much more restricted choice of tools and resources at their disposal. And yet, the use of CAI tools is currently on the rise, as they allow interpreters to prepare assignments ahead of time, provide them with assistance during the actual interpretation and even aid in post-processing (Sandrelli and Jerez, 2007; Costa et al., 2014; Fantinouli and Prandi, 2018). Speech

recognition, in particular, has recently caught the attention of scholars to be used as a central component of CAI tools, either bundled or standalone. In fact, ASR could be considered a major turning point in the growing trend towards digitalisation and technologisation of interpreters' work conditions.

This study offers a comparative analysis of S2T technology for interpreters, and the opportunities ASR opens up as a documentation aid (terminology, background knowledge, domain survey, etc.). Our main aim is to establish the most suitable ASR application for building ad hoc corpora prior to an interpretation assignment. To the best of our knowledge, this is one of the few recent studies to experiment with S2T technology as a CAI tool. The approach we introduce uses, for the first time, the automatic transcription of spoken speeches as an interpreting documentation aid by compiling an ad hoc corpus and extracting candidate terms. Such an approach can lead to more future ASR-oriented interpreting researches.

The paper is organised as follows. Section 2 discusses ASR and S2T within the broader category of CAI tools. Section 3 covers the methodological framework used for data collection and preparation. The comparative analysis and evaluation of ASR applications are described in Sections 3.1 and 3.2 (3.2.1), respectively. Corpus-based term extraction is covered in Section 3.3. Section 3.4 discusses the top-performing ASR system from those analysed in our study. Finally, Section 4 presents our conclusions, the limitations of this study and further research.

2. SPEECH-TO-TEXT TECHNOLOGY AS A COMPONENT OF CAI TOOLS

Motivated by a desire to enhance human-human and human-machine communication, research into ASR and its practical applications has evolved over the past five decades (Yu and Deng, 2015). Generally speaking, ASR has been implemented and integrated into different industries and sectors: voice search, personal digital assistant, gaming, living room interaction systems, and in-vehicle infotainment systems (ibid.). In addition, different categories or applications of ASR started to be used, providing multiple services. For instance, S2T has been deployed to help hearing-impaired people in various settings, like viewing TV programmes, taking educational courses (Stinson et al., 1999), and participating in conferences and awareness campaigns, to name but a few. On the other hand, Text-to-Speech (T2S), which simply refers to the conversion of written text to speech (also termed "synthesis"), has been used to provide more accessibility of written text to visually impaired people and non-native speakers (Quintas, 2017).

As far as the use of ASR in interpreting is concerned, the uptake of such technology has not, till now, been well integrated into the interpreter's workstation. Although significant interest in this technology has arisen in recent years, the state of the art still suffers from a lack of empirical studies dedicated to the use of ASR in interpreting (Cheung and Tianyun, 2018). However, thanks to the improvement in quality of ASR systems, as a result of advancements achieved by deep learning and neural networks as well as growing commercial interest from software companies, the integration of S2T into CAI tools has begun to progress. For instance, Voice-to-Text devices have been integrated into CAI tools to satisfy interpreters' needs in different interpreting contexts and modes (Corpas Pastor, 2018). A wide range of voice dictation and S2T apps have been developed and optimised to be compatible with different operating systems (MacOS, Windows, iOS and Android)

in order to be used for teaching, improving language skills and supporting S2T tasks (Costa, Corpas Pastor and Durán, 2014). Nowadays, the potential of corpus-based ASR for prediction/rendering of untranslated terms by interpreters is being researched at Carnegie Mellon (Vogler et al., 2019).

Furthermore, thanks to the technological advances experienced through Natural Language Processing (NLP) and Artificial Intelligence (AI), Speech-to-Speech (S2S) translation has been one of the most promising attempts of ASR to support human-human communication (Ali and Renales, 2018). S2S translation, being one of the most important applications of ASR and machine translation (MT), has recently been the focus of large projects and software companies: European TC-Star, which dealt with S2S translation of speeches delivered at the European parliament; the DARPA-funded GALE project, which aimed to translate Arabic and Chinese broadcast news into English; and mobile applications developed by Google (Peitz et al., 2011). This technology of S2S translation, also known as Spoken Language Translation (SLT) or Machine Interpretation (MI), conventionally entails three separate components: Automatic Speech Recognition, Machine Translation and Text-to-Speech Synthesis. However, recent efforts have replaced this three-step-software process with a direct S2S translation without relying on an intermediate text representation (Weiss et al., 2017). Unlike the S2S translation, where both the input and output are spoken speech, S2T translation has been used and tested in classroom environments in order enhance the exchange of ideas and open-ended discussion between teachers and students (Blanchard et al., 2015).

We believe that the potential of ASR can satisfy many needs for different interpreting modes (sight, consecutive and simultaneous) and phases (before, during and after interpreting). In that regard, Pöchhacker (2016) highlights the role of speech recognition in supporting court interpreters through instant transcription. He also advocates its potential to reshape the professional practice in general (ibid.). In the same vein, Fantinouli (2016: 50) proposed this technology as "the next step in the evolution of CAI tools". The automatic information extraction of named entities, numbers, etc. in real-time could represent some of its benefits for interpreters during simultaneous interpreting (ibid.). As for consecutive interpreting, ASR could be integrated into CAI tools so that the output may convert the process into sight translation.

In regard to the research (empirical results) about ASR as a CAI tool, Fantinuoli (2017b) introduced ASR as querying system during simultaneous interpreting, establishing several requirements for a successful integration of ASR into a CAI tool, such as being speaker-independent, having the capacity to operate on continuous speech, supporting large-vocabulary recognition, detecting specialised terms, and having high accuracy and speed.

As a further step to boost the use of such technology, Desmet et al. (2018) conducted an experimental study to evaluate the feasibility of using ASR systems (specifically automatic number recognition) to determine whether or not it is helpful for interpreters in-booth. The study concluded that technological support was able to reduce the cognitive loads and improve interpreting quality from 56.5 to 86.5 per cent.

With the aim of examining the usefulness of real-time transcription generated by ASR in-booth, Cheung and Tianyun (2018) carried out a pilot experiment providing the interpreters with the transcription of speeches delivered

in a non-standard accent. The study reported that the fluency score improved when using the transcriptions generated by the ASR during the interpreting process.

3. METHODOLOGY

In order to establish the most suitable ASR application for interpreting-orientated, speech-based term extraction, a six-step methodology was developed for this study. First, an exploratory phase was carried out to evaluate and compare the available and suitable ASR tools by using them to transcribe a set of ten speeches that had been previously collected. For this purpose, the BLEU (Bilingual Evaluation Understudy) tool was used to measure the accuracy of the available software applications. Based on the results obtained from the BLEU score, the most accurate output is selected as a monolingual corpus which consequently will be uploaded into a Terminology Extraction (TE) tool to get a terms list. Our methodology is illustrated below.

3.1 Data collection and preparation

Specialised terminology plays an essential role in various interpreting modes and settings, and it is considered one of the most important concerns for any interpreter during the preparatory phase. Nevertheless, interpreters also need to become familiarised with general terms, either to understand the technical concepts through less specialised material or to render any utterance given by non-specialised speakers. However, terminology management turns into a highly demanding and difficult task, as interpreters are not always provided with the conference program or reference materials beforehand (Gallego Hernández and Tolosa, 2012). For any interpreting assignment, several scenarios should be considered, and different language registers and types of speeches need to be taken into account by the interpreter (Gile, 1990; Setton, 1999; Kalina, 2000), especially during the preparation phase.

For our study, we adopted a "blind interpreting assignment" approach, i.e., a situation in

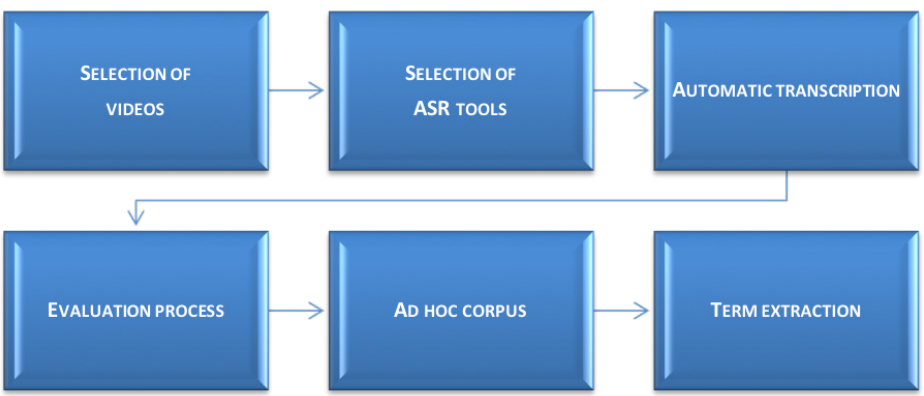

FIGURE 1. *Methodology phases*

which the organiser would provide the interpreter only with the conference topic without giving more details about the speakers' expertise nor their accent, speech type and so on. We opted for blind interpreting in order to include a variety of spoken speeches that should fulfil specific criteria (see below) instead of limiting the search to a given speaker. This enabled us to broaden our research and assess ASR tools with regard to their performance and robustness. In this context, we simulated a preparatory phase (advance preparation) for a blind interpreting assignment on "climate change". We chose this topic because it is one of the most widely debated issues at national and international spheres (Lam et al., 2019). To this end, we selected ten videos according to specific criteria regarding length, degree of difficulty and specialisation, setting, accent and background sound:

- *Length*: The collected videos have a different length, which varies between 04:59 to 48:46 minutes. This longer video was selected to test the consistency of the different ASR engines.
- *Degree of difficulty*: As with any preparation process, and with the purpose to acquire the subject knowledge, a professional interpreter may tend to go from the general material with basic terms, addressed to a layperson in the field, delving, step by step, into more technical information, aimed at specialised individuals. This criterion was included to account for gradual difficulty.
- *Degree of specialisation*: At any event, the specialised language used may vary depending on the speakers' backgrounds and expertise: political, scientific, academic, professional, etc. This criterion is included to provide a clear judgment about the accuracy of each ASR system across general and specialised terms.
- *Setting*: Speeches (delivered at both national and international organisations), training courses and media reports have been selected to have more setting variety.
- *Accent*: American and British accents can be found on the set of selected videos. With this criterion, the performance of each ASR engine is tested in regard to the transcription accuracy across different speakers' accents.
- *Sound background*: Some of the selected speeches are delivered with a musical background, whereas others are not.

The established criteria intend to create a frame to collect specific videos that, from our point of view, could be, on the one hand, useful for the needs of a blind interpreting assignment, and, on the other hand, helpful for testing the performance of the selected ASR tools within different audio-visual contexts. Each one of the aforementioned selection criteria plays an important role when assessing the accuracy, performance and efficiency of the ASR systems analysed.

All selected materials were obtained from official institutions' YouTube channels: Intergovernmental Panel on Climate Change (IPCC), University of British Columbia (UBC), United Nations Framework Convention on Climate Change (UNFCCC) and The Obama White House.

The table below shows the features and characteristics of the videos we collected about climate change.

3.2. ASR tools

This section covers evaluation and comparison of ASR tools with a view to selecting the most appropriate one in the context of corpus building and terminology extraction for advance preparation in interpreting. For reasons of easy access and availability, we have focussed on freeware tools able to perform the automatic transcrip-

Table 1. Videos characteristics and info

Video code	Title	Length/ minutes	Degree of specialisation	Setting	Music background	Accent
V1	*Framing the Climate Change Conversation*	14:25	High	Academic presentation	No	American
V2	*Introduction to Climate Change Impact*	10:10	High	Academic presentation	No	American
V3	*What is Climate Change Mitigation*	09:31	High	Academic presentation	No	American
V4	*Adaptation Strategies*	10:24	High	Academic presentation	No	American
V5	*Adapting to a Changing Climate*	19:33	High	Documentary	Yes	Mixed
V6	*David Cameron's Speech*	04:59	Medium	Official conference	No	British
V7	*English - Climate Change 2014*	11:48	General	Media report	Yes	Mixed
V8	*Fifth Assessment Report*	14:40	General	Media report	Yes	Mixed
V9	*President Obama Speaks on Climate Change*	48:46	Medium	Official conference	No	American
V10	*Prime Minister's Speech on the Environment*	25:23	Medium	Official onference	No	British

tion. To this end, we initially selected eleven free, or at least semi-free, tools that do not require any training or optimising to transcribe the audiovisual material. Tools such as Speechware, Descript, Transcribe, Dragon, Temi, Trint AI, Speechmatics, Spoken Online, Spext, Sonix, Transana, Inqscribe and the like have been discarded, either because they require a commercial licence or offer very limited free minutes.

Once the audiovisual material and ASR were decided, we started to test each one of the eleven tools with the selected ten speeches. During the transcription process, two of them (WEB Speech API demonstration and TalkTyper) stopped functioning continuously, even after several trials and for different videos. For this reason, we discarded both of them and limited our evaluation to the following nine tools: Otter AI, YouTube, IBM´s Watson^{Beta}, Google Docs, SpeechTexter, Speechnotes, TextFromToSpeech, SpeechPal and Dictation.

During the test process, apart from the application function requirements that we had to take into account, we needed to deal with some issues that required technical solutions. For instance, IBM's Watson^{Beta} supports only audio files and specific extensions, some applications provide the transcription only by dictating the speech, whilst others allow file uploading or web link inserting, etc. In the case of the applications that support only speech dictation, the ASR system seemed to be weakened by external background

noise, which was reflected on its accuracy rate. To cope with this issue, the ideal solution was to install a Virtual Audio Cable (VAC[1]) which acts like a physical cable and transfers the audio from the video playing in the background to the ASR microphone without any noise or loss in sound quality. With regards to the file type and extension required for IBM's Watson, we used a third-party application[2] to convert the videos into the required file type and extension. For the applications that only allow file uploading, we relied on RealPlayer to download the speeches from YouTube to get video files that could be uploaded on the transcriber platform.

Finally, ten different transcriptions per utterance were generated by the selected nine tools; i.e. we had at our disposal a total of ninety transcripts. The output of each one of the nine tools was submitted to the evaluation process, which is discussed in the following section. As a result of our research and the transcription process, comparison tables were elaborated (Table 2 and Table 3). Due to format constraints, the results are presented in two tables. Table 2 compares three features: licence type, transcription method and supported languages of each ASR tool. Languages in Table 2 are represented by their international code (cf. ISO 639-1). Table 3 presents what could be deemed as advantages or disadvantages for interpreters when they deal with any ASR tool, such as speaker identification options, the number of supported languages, keyword extraction, punctuation, export output to various formats, click on any word in the transcript to listen to it again, and the required format of the audio/video. For instance, the option of speaker identification could be useful for transcribing videos or audios that contain more than one speaker. Interpreters who work with multiple languages would be interested in familiarising themselves with such ASR tools that provide transcriptions for many languages. Since most ASR systems do not provide punctuation prediction, it is considered a potential feature to be taken into account (Peitz et al., 2011). The ease of exporting the output of speech recognition systems to various formats would be rather practical to meet the requirements of many corpus management applications.

Google Docs is the application that supports the greatest variety of languages and dialects (more than 60), followed by SpeechTexter (44 languages), Speechnotes and Dictation (both of them support over 40 languages and dialects), YouTube (ten languages), IBM's Watson[Bet] and TextFromToSpeech (both of them support nine languages). Both SpeechPal and Otter AI support English only.

In general, most ASR systems provide the output without punctuation marks (Peitz et al., 2011). Therefore, only three tools (Otter AI, IBM's Watson [Beta], SpeechPal) out of the nine analysed transcribe the speech providing punctuation marks. As indicated in Table 2, there are three ways to get the speech transcribed by an ASR system: simultaneous dictation, video file upload or by inserting the video's web link. Most ASR applications include simultaneous dictation (seven out of nine).

3.2.1. Evaluation and results

To obtain quality measurement on the transcriptions generated by ASR applications, the output (hypothesis text) of each ASR tool is compared to a human transcription (reference text) of input speech (González et al., 2011). Both Word Error Rate (WER) and Word Accuracy (WAcc) are stan-

[1] The VAC can be download at: https://www.vb-audio.com/Cable/

[2] In this case we used My MP4 to MP3 Converter, which is a software application developed by Microsoft to convert an mp4 file into an mp3 file. https://www.microsoft.com/en-us/p/my-mp4-to-mp3-converter/9nblggh6j02v?activetab=pivot%3Aoverviewtab

dard measures widely used to evaluate the quality of ASR systems (ibid). The BLEU metric has been used to score reference texts and hypothesis texts for written-language machine translation (Papineni et al., 2002), the interaction of ASR with machine translation in a speech translation system (Condon et al., 2008; He, Deng and Acero, 2011; Deiru et al., 2019; Khandelwal, 2020) as well as any output text for a suite of natural language processing tasks (Brownlee, 2017).

- *WER:* It measures the performance of an ASR, comparing a reference to a hypothesis:

$$WER = \frac{S + D + I}{N}$$

(S = number of substitutions; D = number of eliminations; I = number of insertions; N = number of words in the reference)

- *WAcc:* It measures the total number of correct words in the hypothesis text in relation to the total number of words in the reference text:

$$WAcc = \frac{N - D - S - I}{N}$$

- *BLEU score:* BLEU score can range from 0 to 1, where higher scores indicate closer matches to the human transcription, i.e., the closer an ASR output is to human transcription, the better it is.

Table 2. A comparison of the ASR applications: functionalities and languages supported

TOOL	LICENCE TYPE	IMPORT VIDEO/ AUDIO (S2T)	VR/ DICTATION	INSERT WEB LINK	LANGUAGES
Otter AI	Free 600 mins/ Month	√	√	x	EN
YouTube	Free	√	x	x	EN, ES, FR, DE, IT, JA, KO, NL PT, RU
IBM's Watson Beta	Free	√ (only audio)	√	x	AR, EN, ES, FR, PT, PT_Brazil, JA, KO, DE, ZH
Google Docs	Free	x	√	x	AR, DE, EN, ES, FR, IT, PT, RO, RU, ZH, etc.
Speech Texter	Free	x	√	x	DE, ES, FA, GR, HE, HI, HU, IT, JA, PO, PT, RO, RU, SU, TK, UK, UR, etc.
Speech notes	Free	x	√	x	AR, BU, DE, EN, ES, FR, IT, NL, PT, RO, TK, etc.
Text-FromTo Speech	Free	√	√	x	AR, EN, ES FR, IT, JA, NL, RU, UK,
SpeechPal	Free 120 mins	√	x	√	EN
Dictation	Free	x	√	x	AR, EN, ES, FR, IT, NL, TK, etc.

Table 3. A comparison of the ASR applications: pros and cons

Tool	pros	cons
Otter AI	Speaker identification. Punctuation. Keywords. Export output to various formats (.txt, .pdf, .srt), web link and copy to clipboard. Click on any word in the transcript to listen to it again.	Supports English only.
YouTube	Supports ten languages. Click on any word in the transcript to listen to it again.	No punctuation.
IBM's Watson Beta	Supports nine languages. Speaker identification. Click on any word in the transcript to listen to it again.	File format limitation: .mp3, .mpeg, .way, .flac, or .opus only.
Google Docs	Supports more than 60 languages and dialects.	No punctuation. Disconnection.
SpeechTexter	Supports 44 languages. Export output to various formats: .txt, .doc and copy to clipboard.	No punctuation.
Speechnotes	Supports 44 languages and dialects. Export output to various formats: .txt, .doc, upload to Google drive and copy to clipboard.	No punctuation.
TextFromTo Speech	Supports nine languages. Export output to various formats: .txt, .doc, copy to clipboard or email the dictated text.	No punctuation.
SpeechPal	Punctuation. Export output to .txt or email the dictated text. Click on any word in the transcript to listen to it again.	Supports English only.
Dictation	Supports more than 40 languages and dialects. Export output to txt or email the dictated text.	No punctuation.

Generally, the preference to decide on the ASR evaluation metric depends on the ASR system function, whether it is query-oriented, i.e., to recognise named entities and identify specific terms, or to perform an S2S translation, etc. For our study, which focuses on automatic transcription, the evaluation is performed using the BLEU score (evaluating the transcription of ten videos) and WER (evaluating the transcription of five videos). A gold-standard transcription was manually prepared for each video in order to be used as a reference text and to be compared against the automatic transcription obtained by each ASR tool. Once all the gold-standard transcriptions were ready, we started the evaluation task. For this purpose, a Python library NLTK[3] was used in order to obtain the BLEU scores for the generated transcripts. Figure 2 (below) illustrates the performance and accuracy of each tool using BLEU with the ten selected speeches. Otter AI scored the highest performance, followed by YouTube and SpeechPal.

In order to confirm and validate the BLEU results, we performed an assessment for the five videos using the WER measure (V1-V5). WER (see Table 4) confirmed and provided the same results of BLEU score, Otter AI being the more ac-

[3] The Python library NLTK is available at https://www.nltk.org/api/nltk.translate.html.

curate tool followed by YouTube and SpeechPal.

At this stage, the evaluation was completed for the nine ASR tools relying on both BLEU score and WER metrics. The next phase was the exploitation of the transcripts obtained by Otter AI through the use of a corpus management application.

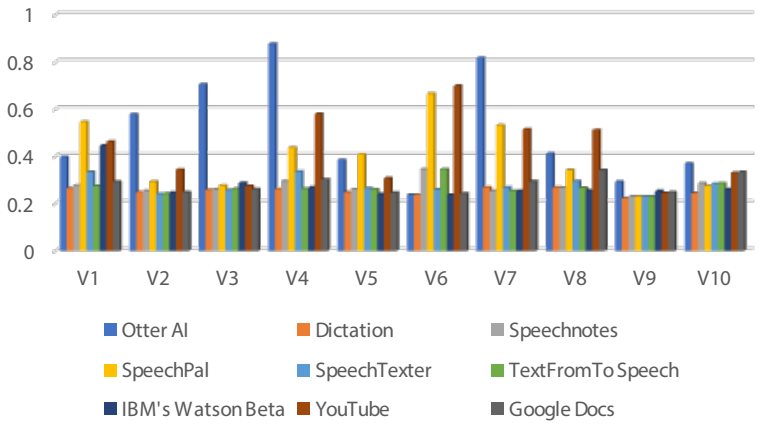

FIGURE 2. *BLEU score results for ASR applications across the all videos (V1-V10)*

Table 4. WER results for ASR tools performance for five videos (V1-V5)

	WER				
Tool	**V1**	**V2**	**V3**	**V4**	**V5**
Otter AI	0.01	0.005	0.004	0.001	0.008
Dictation	0.3455	0.275	0.261	0.297	0.354
Speechnotes	0.1645	0.151	0.151	0.141	0.56
SpeechPal	0.0225	0.026	0.038	0.024	0.054
SpeechTexter	0.2135	0.368	0.198	0.276	0.416
TextFromToSpeech	0.1645	0.208	0.191	0.652	0.242
IBM ´s Watson [Beta]	0.044	0.072	0.063	0.078	0.122
YouTube	0.0325	0.026	0.029	0.026	0.031
Google Docs	0.183	0.108	0.12	0.123	0.295

3.3. Ad hoc *corpus and term extraction*

In this section we suggest exploiting the ASR outcome as an ad hoc corpus. Corpus Driven Interpreters Preparation (CDIP) and what is also called corpus-based terminology preparation can improve the interpreter's performance on specialised topics (Fantinouli, 2006; Bale, 2013; Xu, 2018; Pérez-Pérez, 2018; Gallego Hernández and Tolosa, 2012; Sánchez Ramos, 2017; and Section 1 of this paper). The state of the art shows that ad hoc corpora of written texts have been already deployed as an effective solution to manage terminology and acquire the necessary knowledge for any specialised topic.

In interpreting, TE is used to "identify a list of monolingual specialised terms and phrases from the collected corpus that can be used by the interpreter to create a conference glossary as well as to start the learning process" (Fantinuoli, 2017a: 33). TE is considered an important contribution and effective solution to acquiring expert knowledge in any field. It also makes it easier to retrieve information for a given interpreting assignment.

Based on the results of BLEU and WER measurements, Otter AI comes out as the most effective ASR system against the others, which means that its output is the most accurate one. The transcripts generated by this tool, like the other tools evaluated in this study, provided us with a monolingual native-language corpus comprised from public speeches in a specialised subject field, in this case *climate change*. This output, considered as an exploitable "raw material", is uploaded into a piece of corpus management software in order to accomplish two objectives: (a) manage the ten transcripts obtained by the ASR application; and (b) compile and download a .txt file corpus comprised from the ten transcripts to be used later for the automatic term extraction.

In this study we used Sketch Engine[4] to manage the obtained transcripts. This software is a comprehensive suite of tools that enables corpus management, text analysis, concordancing, keyword and n-gram extraction, as well as other functionalities (but not term candidates). Figure 3 illustrates the concordance function of Sketch Engine operating on our *Climate Change* corpus.

We used Sketch Engine to generate our corpus as .txt files from various documents. The resulting ad hoc corpus was subsequently uploaded into an automatic term extractor. Experiments and studies have shown that the use of automatic term extraction, as part of the preparation phase, can improve the performance of interpreters during the simultaneous interpreting (Fantinouli, 2006; Gallego Hernández and Tolosa, 2012; Xu, 2015).

There is a wide range of tools[5] for both monolingual and bilingual term extraction, based on linguistic, statistical or even hybrid approaches, in either open-source or commercial software: TerMine (Frantzi et al., 2000), YATE (Vivaldi, 2001), TermoStat (Drouin, 2003), Terminus[6], Linguoc LexTerm (Oliver et al., 2007), TermSuite (Daille, 2012), ProTermino (Durán Muñoz et al., 2015), TermStar[7], etc.

For this study we have used the Terminology Extraction Suite (TES)[8] (Oliver and Vázquez, 2007) to get a list of the candidate terms. TES is written in Perl, can be run on Linux, Windows and Mac and uses a statistics-based approach to automatically extract terminology. It can ex-

[4] Available at: https://www.sketchengine.eu/.

[5] See Zaretskaya, Corpas Pastor and Seghiri (2015).

[6] Available at: http://terminus.iula.upf.edu/cgi-bin/terminus2.0/terminus.pl?lInt=En.

[7] Available at: https://www.star-group.net/tr/products/termstar.html.

[8] This application can be downloaded from: https://sourceforge.net/projects/terminology-extraction-suite/.

tract candidate terms from monolingual and bilingual corpora as well. Once the candidate terms are generated (TES-Wizard), and since the Terminology Extraction Suite allows multiple selection of terms to be eliminated, manual revision can be applied to discard the words (i.e., the general words that are not typically related to a specialised field) and keep the terms (which represent a specific concept linked to a specific field or discipline), so that a monolingual glossary can be compiled (see Fig. 4).

Finally, the TES-editor enables selection of relevant candidate terms to be exported as a list of terms (glossary) in various formats.[9]

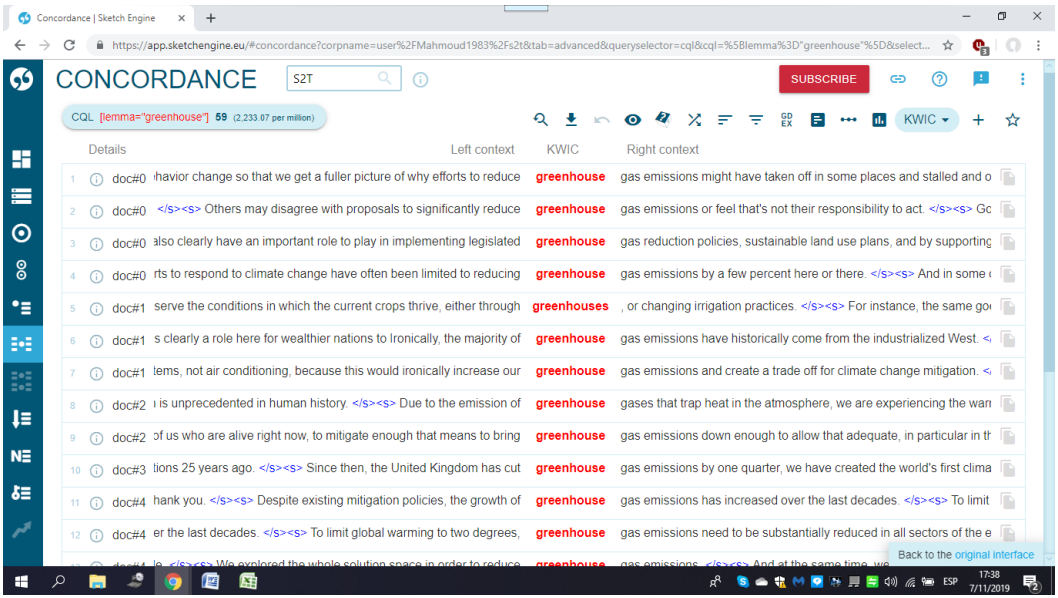

FIGURE 3. *KWIC concordance for greenhouse*

[9] TES-editor can also calculate translation equivalents provided the suite is run on a parallel corpus.

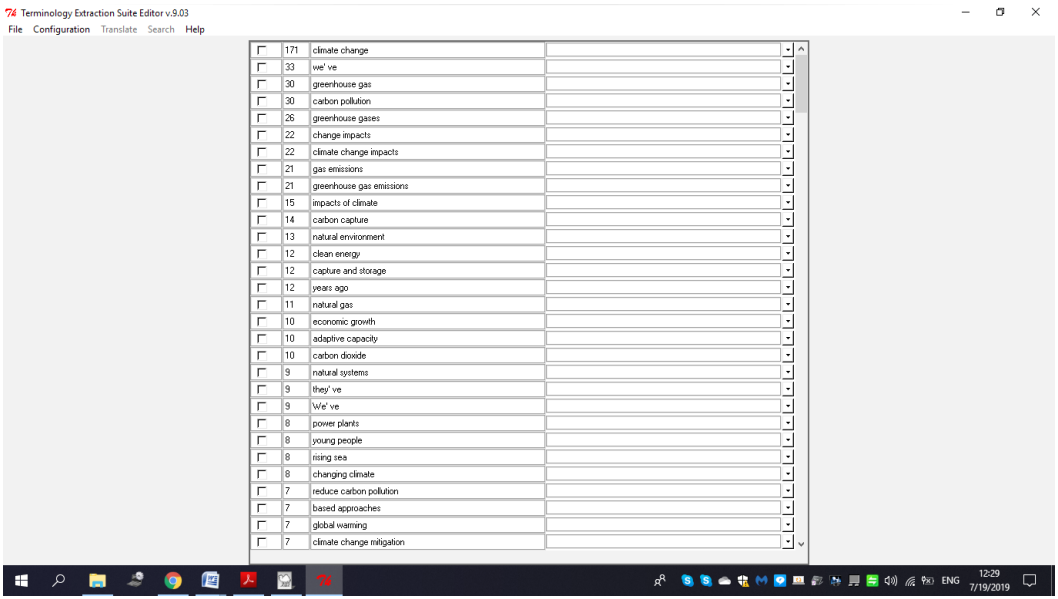

FIGURE 4. *Candidate terms retrieved by TES (v9.03)*

3.4. Choice of ASR application

The previous sections offer a global picture of the advantages and drawbacks of the ASR systems analysed. Some tools required a specific file format and a certain amount of previous use/training to get the most of them, while others were ready to be used without any requirements. It is worth mentioning that we logged the start and end time for every transcription, which demonstrated that some tools were able to perform the transcription in less time than the original video duration. For instance, video 9 is 48:46 minutes long and was transcribed in only 21 minutes by Otter AI.

The evaluation showed that Otter AI had the best performance compared to the rest of the tools in this study. Apart from the high score, Otter AI also offers the keyword extraction feature (see Figure 5), which is a valuable function for any translator or interpreter. Although the mentioned feature is not as sophisticated as many other automatic terminology extraction tools (TES, for instance), it is still an added value for the application. For instance, many term extraction tools can be pre-programmed, depending on their extraction approach (statistic, linguistic or hybrid) by the user selecting specific criteria, such as frequency values (setting the minimum frequency that a lexical item must have in order to be listed as a candidate term), patterns or measurements of association between the elements of a multiword unit (e.g. noun+preposition+noun, adjective+noun, etc.), *n*-grams limit, list of stop words, etc. All these parameters/functionalities have not been incorporated into Otter AI so far. Despite the high score of Otter AI, the language limitation, since it only supports English, and not being fully free

software, can be considered drawbacks for our scope of work.

YouTube demonstrated a good accuracy rate, being the second-best tool after Otter AI, with the advantage of supporting ten languages. Like most of all state-of-the-art ASR systems, YouTube recognises sequences of words but does not provide punctuation marks. To make the most of this tool, considering that it supports various languages, we suggest the use of any punctuation prediction method[10] to cope with such issue and enrich the ASR system output with punctuation marks. In addition, considering the quality of transcripts and the number of languages supported by YouTube, an interesting possibility could also be to use a TE system to extract terms automatically from the corpus of transcripts generated by YouTube, perform human and automatic evaluation of YouTube transcripts and Otter AI transcripts, and compare results of term extraction for both.

Finally, it is worth noting that the performance of most of the tools was not consistent enough across all videos. We assume that the tools' performance inconsistency was due to the characteristic variation of the selected videos: music background, accent, duration, degree of language difficulty and/or specialisation, etc.

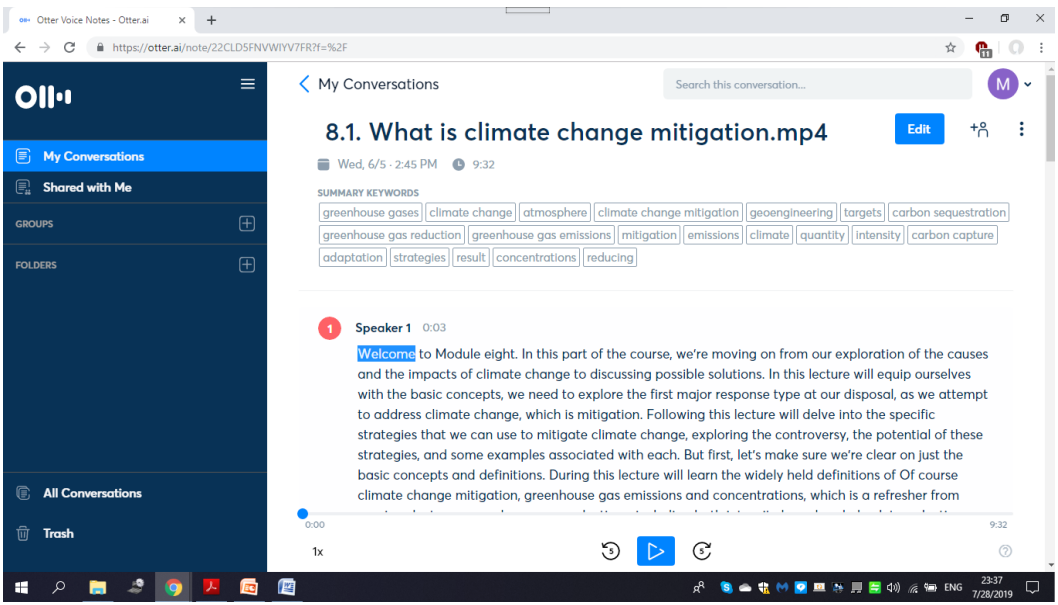

FIGURE 5. *Keyword extraction in Otter AI*

[10] For more information with regard to punctuation prediction approaches, see Peitz, Freitag and Mauser (2011) or Kolár and Lamel (2012).

4. CONCLUSION

In this study, we introduced a new approach to applying one of the ASR benefits, which is S2T technology, and further process of its output to satisfy specific interpreter preparatory needs. To this end, a six-step methodology has been used. By means of a comparative study on nine ASR tools, and using data from the burgeoning field of climate change discourse, we have been able to establish the most accurate ASR tool for ad hoc corpus compilation and term extraction from video recordings of speeches by means of S2T technology. This approach provides interpreters with a novel technological solution for advance preparation in the documentation phase. Corpus compilation and terminology extraction are only two examples of the benefits we can obtain from this technology. Although ASR technology is still far from perfect, automatic transcriptions reveal a very valuable resource for interpreters, with very promising results so far.

Our study presents several limitations. We only used one language (English) to evaluate the performance of the ASR tools, which means that our results might not be extrapolatable to other languages. We also opted for a small-size corpus (total of 170 minutes and 23, 757 words), that was fit for the purpose (considering the aims of our study on the feasibility of ASR transcription, and to confirm or disprove our hypothesis). Further research is needed with more languages and large-scale corpora of recorded speeches. In addition, we only examined the individual performance of ASR systems for corpus compilation and term extraction purposes, but we did not take into account combined results for term extraction, such as uploading the corpus generated by YouTube (or any other ASR system for that matter), using a TE system, and then performing human and automatic evaluation of YouTube transcripts and Otter AI transcripts, for comparison.

Finally, we plan to explore the potential of ASR technology in depth in order to develop an integrated and multifunctional CAI tool that will be able to transcribe speech, compile and manage spoken corpora, extract terminology and multiword units and perform speech queries. Interpreting is rapidly heading towards digitalisation and technologisation. In this new context, interpreters should be equipped with appropriate tools and resources before they find themselves stuck in the technology whirlpool.

DECLARATION

Authors of the present paper do not have any relationship or commercial interest with any of the mentioned software companies. The aim of this study is merely for academic research purposes.

Acknowledgement

The research reported in this study has been carried out within the framework of research projects VIP (Ref. no. FFI2016-75831-P, 2017-2020), TRIAJE (Ref. UMA18-FEDERJA-067) and MI4ALL (CEI, Andalucía Tech). The authors wish to thank the editor and two anonymous reviewers for their useful comments and suggestions.

REFERENCES

Ali, Ahmed, and Steve Renals (2018): "Word Error Rate Estimation for Speech Recognition: e-WER", *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Short Papers), Melbourne, Australia, 20–24.

Arce Romeral, Lorena and Míriam Seghiri (2018): "Booth-friendly term extraction methodology based on parallel corpora for training medical inter-

preters", *Current Trends in Translation Teaching and Learning E*, 5, 1-46.

Bale, Richard (2013): "Undergraduate Consecutive Interpreting and Lexical Knowledge", *The Interpreter and Translator Trainer*, 7/1, 27-50.

Blanchard, Nathaniel, Michael Brady, Andrew M. Olney, Marci Glaus, Xiaovi Sun, Martin Nystrand, Borhan Samei, Sean Kelly, and Sidney D'Mello (2015): "A study of automatic speech recognition in noisy classroom environments for automated dialogue analysis", *International Conference on Artificial Intelligence in Education*, Springer, Cham, 23–33.

Brownlee, Johnson (2017): "A Gentle Introduction to Calculating the BLEU Score for Text in Python", *Machine Learning Mastery*, 20th Nov <https://machinelearningmastery.com/> [Accessed: 04-VI-2020].

Cheung, Andrew KF, and Li Tianyun (2018): "Automatic speech recognition in simultaneous interpreting: A new approach to computer-aided interpreting", *Proceedings of Ewha Research Institute for Translation Studies International Conference*, At Ewha Womans University.

Condon, Sherri L., Jon Phillips, Christy Doran, John S. Aberdeen, Dan Parvaz, Beatrice T. Oshika, Gregory A. Standers, and Craig Schlenoff (2008): "Applying Automated Metrics to Speech Translation Dialogs", *Proceedings of LREC-May 2008*.

Corpas Pastor, Gloria and Isabel Durán-Muñoz (eds.) (2017): *Trends in E-tools and Resources for Translators and Interpreters*, Leiden, Holland: Brill | Rodopi. https://doi.org/10.1163/9789004351790.

Corpas Pastor, Gloria (2018): "Tools for Interpreters: The Challenges that Lie Ahead", *Current Trends in Translation Teaching and Learning E*, 5, 157-182.

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz (2014): "Technology-assisted Interpreting", *MultiLingual,* 143/25, 27-32.

Daille, Béatrice (2012): "Building bilingual terminologies from comparable corpora: The TTC TermSuite", *Proceeding of 5th Workshop on Building and Using Comparable Corpora at LREC 2012*.

Deng, Li, and Douglas O'Shaughnessy (2003): *Speech Processing: A Dynamic and Optimization-Oriented Approach*, New York: Marcel Dekker Inc.

Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak (2019): "Survey on evaluation methods for dialogue systems", *arXiv preprint arXiv:1905.04071*, 1, 1-62 <https://arxiv.org/abs/1905.04071> [Accessed: 04-VI-2020].

Desmet, Bart, Mieke Vandierendonck, and Bart Defrancq (2018): "Simultaneous interpretation of numbers and the impact of technological support", in Claudio Fantinuoli (ed.), *Interpreting and technology,* Berlin: Language Science Press, 13–27, doi:10.5281/zenodo.1493291.

Drouin, Patrick (2003): "Term extraction using nontechnical corpora as a point of leverage", *Terminology*, 9/1, 99-115.

Durán Muñoz, Isabel, Gloria Corpas Pastor, Le An Ha, and Ruslan Mitkov (2015): "Introducing ProTermino: A New Tool Aimed at Translators and Terminologists", *Traducimos desde el Sur, Actas del VI Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación. Las Palmas de Gran Canaria, 23-25 de enero de 2013*, Las Palmas de Gran Canaria: Universidad de Las Palmas de Gran Canaria, Servicio de Publicaciones y Difusión Científica, 623-638.

Fantinuoli, Claudio (2006): "Specialized Corpora from the Web for Simultaneous Interpreters", in Marco Baroni and Silvia Bernardini (eds.), *Wacky! Working papers on the Web as Corpus*, Bologna: GEDIT, 173-190.

Fantinuoli, Claudio (2016): "InterpretBank: Redefining computer-assisted interpreting tools", *Proceedings of the 38th Conference Translating and the Computer*, 42–52.

Fantinuoli, Claudio (2017a): "Computer-assisted preparation in conference interpreting", *Translation & Interpreting* 9/2, 24-37.

Fantinuoli, Claudio (2017b): "Speech recognition in the interpreter workstation", *Proceedings of the Translating and the Computer 39 Conference*, London: Editions Tradulex, 367–377.

Fantinuoli, Claudio and Bianca Prandi (2018): "Teaching information and communication technolo-

gies: a proposal for the interpreting classroom", *Transkom* 11/2, 162-182.

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mima (2000): "Automatic recognition of multi-word terms", *International Journal of Digital Libraries*, 3/2, 117-132.

Gallego Hernández, Daniel and Miguel Tolosa (2012): "Terminología bilingüe y documentación ad hoc para intérpretes de conferencias. Una aproximación metodológica basada en corpus", *Estudios de Traducción*, 2, 33-46.

Gile, Daniel (1990): "Scientific research vs. personal theories in the investigation of interpretation", in Laura Gran and Christopher Taylor (eds.), *Aspects of applied and experimental research on conference interpretation*, Udine: Campanotto Editore, 28-41.

González, María, Julián Moreno, José Luis Martínez, and Paloma Martínez (2011): "An illustrated methodology for evaluating ASR systems", *International Workshop on Adaptive Multimedia Retrieval*, Berlin: Springer, 33-42.

He, Xiaodong, Li Deng, and Alex Acero (2011): "Why word error rate is not a good metric for speech recognizer training for the speech translation task?", *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 5632-5635.

Kalina, Sylvia (2000): "Interpreting competences as a basis and a goal for teaching", *The Interpreters' Newsletter*, 10, 3–32.

Khandelwal, Renu (2020): "Blue- Bilingual Evaluation Understudy: A step by setp approach to understanding BLEU, a metric to understand the effectiveness of Machine Translation (MT)", *Towards Data Science,* 25th Jan <https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1> [Accessed: 04-VI-2020].

Kolár, Jáchym and Lori Lamel (2012): "Development and Evaluation of Automatic Punctuation for French and English Speech-to-Text", *Thirteenth Annual Conference of the International Speech Communication Association*.

Lam, Adriane R., Jennifer E. Bauer, Susanna Fraass, Sarah Sheffield, Maggie R. Limbeck, Rose M. Borden, Megan E. Thompson-Munson (2019): "Time Scavengers: An Educational Website to Communicate Climate Change and Evolutionary Theory to the Public through Blogs, Web Pages, and Social Media Platforms", *Journal of STEM Outreach*, 7/2, 1-8.

Oliver, Antoni, and Mercè Vázquez (2007): "A Free Terminology Extraction Suite", *Proceeding of Translating and the Computer 29th Conference*, London.

Oliver, Antoni, Mercè Vázquez, and Joaquim Moré (2007): "Linguoc LexTerm: una herramienta de extracción automática de terminología gratuita", *Translation Journal*, 4/11.

Peitz, Stephan, Markus Freitag, Arne Mauser, and Hermann Ney (2011): "Modeling Punctuation Prediction as Machine Translation", *International Workshop on Spoken Language Translation (IWSLT)*, 238-245.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002): "BLEU: A Method for Automatic Evaluation of Machine Translation", *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 311-318.

Pérez-Pérez, Pablo (2018): "The Use of a Corpus Management Tool for the Preparation of Interpreting Assignments: A Case Study", *The International Journal for Translation and Interpreting Research*, 10/1 137-151.

Pöchhacker, Franz (2016): *Introducing Interpreting Studies*, Abingdon: Routledge, 2nd edition.

Quintas, Laura Cacheiro (2017): "Towards a Hybrid Intralinguistic Subtitling Tool: Miro Translate", in J. Esteves-Ferreira, J. Macan, R. Mitkov and O.-M. Stefanov (eds.), *Proceedings of the 39th Conference Translating and the Computer*. ASLING, London, UK, November 16-17, 2017, Geneva: Tradulex, 01-06.

Sánchez Ramos, María del Mar (2017): "Interpretación sanitaria y herramientas informáticas de traducción: los sistemas de gestión de corpus", *Panace@*, 18/46, 133-141.

Sandrelli, Annalisa, and Jesus De Manuel Jerez (2007): "The Impact of Information and Communication Technology on Interpreter Training: State-of-the-Art and Future Prospects", *The Interpreter and Translator Trainer* 1/2, 269-303, doi:10.1080/1750399X.2007.10798761.

Setton, Robin (1999): *Simultaneous interpretation: A cognitive-pragmatic analysis,* (vol. 28), Amsterdam: John Benjamins Publishing.

Stinson, Micheal S., Sandy Eisenberg, Christy Horn, Judy Larson, Harry Levitt, and Ross Stuckless (1999): "Real-time speech-to-text services", in Ross Stuckless (ed.), *Reports of the National Task Force on Quality Services in Postsecondary Education of Deaf and Hard of Hearing Students.* Rochester, NY: Northeast Technical Assistance Center, Rochester Institute of Technology.

Vivaldi, Jorge (2001): *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas*, Ph.D dissertation, Barcelona: Universidad Politécnica de Catalunya.

Vogler, Nikolai, Craig Stewart, and Graham Neubig (2019): "Lost in Interpretation: Predicting Untranslated Terminology in Simultaneous Interpretation", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* Association for Computational Linguistics, 109–118, doi: 10.18653/v1/N19-1010.

Weiss, Ron J., Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen (2017): "Sequence-to-Sequence Models Can Directly Translate Foreign Speech", *Proceeding of Interspeech*, doi: 10.21437/Interspeech, 2017-503.

Xu, Ran (2015): *Terminology preparation for simultaneous interpreters*, Ph.D dissertation, Leeds: University of Leeds.

Xu, Ran (2018): "Corpus-based terminological preparation for simultaneous interpreting", *Interpreting*, 20/1, 33–62.

Yu, Dong, and Li Deng. (2015): *Automatic speech recognition: A deep learning approach*. London: Springer.

Zaretskaya, Anna, Gloria Corpas Pastor, and Miriam Seghiri (2015): "Translators' requirements for translation technologies: a user survey", *New Horizons in Translation and Interpreting Studies* (Full papers), Geneva, Switzerland: Tradulex, 247–254.