Exploring automatic methods for the construction of Multimodal Interpreting Corpora. How to transcribe linguistic information and identify paralinguistic properties?

Across Languages and Cultures, 2024, 25 (1)

Xiaoman Wang; Binhua Wang
Centre for Translation Studies, University of Leeds, United Kingdom

Xiaoman Wang ORCID id: https://orcid.org/0000-0001-5863-5517
Binhua Wang ORCID id: https://orcid.org/0000-0003-2404-5214
Corresponding author: Binhua Wang, Email address: b.h.w.wang@leeds.ac.uk

Abstract

In corpus-based interpreting studies, typical challenges exist in the time-consuming and labour-intensive nature reproducing transcribing spoken data and in identifying prosodic properties. This paper addresses these challenges by exploring methods for the automatic compilation of multimodal interpreting corpora, with a focus on English/Chinese Consecutive Interpreting. The results show that: 1) automatic transcription can achieve an accuracy rate of 95.3% in transcribing consecutive interpretations; 2) prosodic properties related to filled pauses, unfilled pauses, articulation rate, and mispronounced words can be automatically extracted using our rule-based programming; 3) mispronounced words can be effectively identified by employing Confidence Measure, with any word having a Confidence Measure lower than 0.321 considered as mispronounced; 4) automatic alignment can be achieved through the utilisation of automatic segmentation, sentence embedding, and alignment techniques. This study contributes to interpreting studies by broadening the empirical understanding of orality, enabling multimodal analyses of interpreting products, and providing a new methodological solution for the construction and utilisation of multimodal interpreting corpora. It also has implications in exploring applicability of new technologies in interpreting studies.

Keywords: multimodal interpreting corpus; multi-layer model; automatic extraction of paralinguistic features; disfluency; mispronounced words; automatic alignment

1. Introduction

Corpus linguistics has been applied to the field of interpreting studies for various purposes, including research and education, over the past two decades. Corpus-based Interpreting Studies (CIS) has evolved from its origins as an offshoot of corpus-based translation studies (CTS) (Shlesinger, 1998). It has transitioned from what was once described as a "cottage industry" (Setton, 2011, p 34) to become one of the fastest-growing fields, leveraging the potential of "Web 2.0 and collaborative work" (Bendazzoli, 2018, p 13). This transformation has been made possible by technological advancements in the development of tools for corpus construction, annotation, and analysis. These advancements have given rise to electronic, machine-readable corpora, making it more convenient to provide access to these resources for the entire academic community. Furthermore, they enable computer-assisted inquiries for more in-depth analysis. Notably, research institutes, companies in Natural Language Processing (NLP) and speech technology can utilise these corpora for training NLP models.

Despite the advantages and popularity of constructing interpreting corpora, the creation of multimodal interpreting corpora is widely recognised as a time-consuming endeavour. The process of orthographical transcribing verbal output is particularly labour-intensive, necessitating extensive manual efforts to capture concomitant paralinguistic elements, including prosodic and temporal information, and to align the transcribed source text with the target text. Consequently, the foundational phases of corpus construction significantly influence the size of the interpreting corpus and analytical approaches (Bendazzoli, 2018; Bernardini et al., 2018).This, in turn, limits the potential for automating the extraction of substantial quantities of phenomena for the interrogation of comprehensive language resource and robust analytical outcomes (Falbo, 2018; Russo et al., 2018).

To address this disparity, corpus linguists can enhance efficiency in various stages of corpus construction by leveraging software tools for the creation and analysis of spoken language corpora, thus reducing reliance on entirely manual annotation and transcription. Specifically, one approach to minimize the need for extensive manual labour involves utilising software such as EXMARaLDA (Extensible Markup Language for Discourse Annotation) (Schmidt & Wörner, 2009). This software enables the creation of time-aligned transcription across multiple tiers and allows for the addition of manual annotations across a customizable number of tiers. However, it's noteworthy that there is limited existing literature that explores the transcription of linguistic information and identification of paralinguistic properties in a batch process, as well as the alignment of source and target texts at the sentence level with minimal manual intervention.

Within this context, this study explores the uncharted territory of automatically constructing multimodal interpreting corpora, integrating paralinguistic information into the analysis of interpreters' verbal output. As a significant result of this exploratory methodology, multimodal English/Chinese consecutive interpreting corpora are generated for use by interpreter trainers. This study provides a detailed account of the methodology employed to facilitate computer-aided quantitative analysis of multimodal features. Additionally, the streamlined corpus design method is expected to serve as a technological catalyst, enabling further exploration by future researchers, whether with the same corpora or newly created ones.

2. Literature Review

2.1 Multimodality

The development of social semiotics gained momentum with Halliday's introduction of the term and his exploration of the social dimensions of meaning. He examined how human processes of signification and interpretation shape both individuals and societies (Halliday, 2014). Building on Halliday's ideas, Gunther Kress expanded the scope of social semiotics, moving beyond its linguistic origins to investigate the underlying principles of multimodal communication and develop the theory of multimodality. The concept of multimodality is firmly rooted in the field of social semiotics, focusing on the creation of meaning within social practices. This concept is best encapsulated by the idea of "modes", as defined by Kress and Van Leeuwen: "the use of several semiotic modes in the design of a semiotic product or event" (Kress and Van Leeuwen, 2001, p 20). In a world conceived through a multimodal lens, a mode represents a socially shaped and culturally provided semiotic resource for meaning-making. Examples of modes used in representation and communication include images, writing, layout, music, gesture, speech, moving images, soundtracks, and 3D objects (Kress, 2009). Each of these modes is equally significant in representation and communication, as they operate simultaneously and possess the potential to convey meaning (Kress, 2009).

Interpreting, as a multimodal activity, engages with the intricate semiotics of human behaviour unfolded in the audio-visual reality where language processing comprises auditory perception, oral production and bodily activities. In essence, the multimodality of interpreting is assumed to encompass both verbal and nonverbal sign systems. Poyatos (2002) has provided a comprehensive model for this entire system, describing it as a sign-conveying verbal and nonverbal system that manifests in various situations involving visual and/or acoustic co-presence. According to Poyatos (2002), the audible system comprises verbal language, paralanguage, quasi-paralinguistic sounds emitted through audible kinesics and silence. In contrast, the visible system encompasses elements such as stills, kinesics, as well as visual chemical and dermal systems such as tear (Poyatos, 2002).

Hence, the development of interpreting corpora should not be limited solely to textual materials. Interpreting corpora, as defined, encompass collections of texts, specifically transcriptions of spoken or signed recordings of interpreter-mediated events (Bendazzoli et al., 2018). In the construction of interpreting corpora, it is essential to incorporate paraverbal and nonverbal resources, as they play a fundamental role in meaning-making processes in both direct and mediated communication, as highlighted by Bendazzoli et al. (2020). Multimodal corpora are expected to include transcriptions that encompass not only linguistic elements, such as conventional spelling transcriptions but also paralinguistic properties such as filled and unfilled pauses, mispronounced words, and delivery speed.

2.2 Representative interpreting corpora and their modalities

Over the last two decades, the interpreting community has witnessed the evolution of corpus construction, transitioning from manual monolingual linguistic annotation to the development of fully machine-readable multimodal corpora using automatic or semi-automatic methods. The advent of electronic interpreting corpora has played a crucial role in validating various hypotheses and theories.

Several of these corpora are notably substantial in size when compared to general reference corpora (Bendazzoli, 2018), enabling the extraction of relevant occurrences and

patterns. The majority of these corpora comprise both consecutive and simultaneous interpreting data sourced from conferences, with data collected from various contexts, including European Parliament multilingual plenaries and other conferences. Notably, the European Parliament data serves as the foundation for the creation of large-scale corpora, such as the European Parliament Interpretation Corpus (EPIC) (Russo et al., 2012), which comprises approximately 180,000 words of source texts and their equivalents. Additionally, the European Parliament Interpreting Corpus-Ghent (EPICG) (Bernardini et al., 2018) encompasses around 250,000 words, while the European Parliament Translation and Interpreting Corpus (EPTIC) surpasses 400,000 words (Bernardini et al., 2018). Within the same realm of interpretative contexts, the Directionality in Simultaneous Interpreting Corpus (DIRSIC) (Bendazzoli, 2012) and the CoSI-corpus (House et al., 2012) provide valuable resources for multiple research inquiries. Football press conferences have also been tapped as a data source for conference interpreting, giving rise to the Corpus Football in Europe (FOOTIE) (Sandrelli, 2012). In the domain of dialogue interpreting, medical consultations and court proceedings have provided valuable data for corpora development, with notable examples including the DiK-corpus (Bührig et al., 2012), AIM corpus (Gavioli, 2015), HCIQ.1415 (Dal Fovo, 2018), and the Corpus of Italian judicial hearings (Biagini, 2012). Additionally, the Corpus of Television Interpreting (CorIT) (Falbo, 2012) stands out as the largest multilingual television interpreting corpus, encompassing interpreting performances from over 1200 interpreters, both in conference and dialogue settings.

Within the realm of large-scale interpreting corpora, various interpreting modalities have been explored, particularly in the context of conference interpreting. For instance, Collados Aís et al. (2004) introduced ECIS (Quality Evaluation in Simultaneous Interpreting), which places emphasis on non-verbal and prosodic features. In the case of EPIC, the corpus incorporates not only linguistic elements but also paralinguistic features, including truncated or mispronounced words, as well as filled and unfilled pauses. This comprehensive approach enables researchers to extract two distinct categories of disfluencies present in spoken language. EPIC encompasses a wide range of linguistic aspects, including lexical patterns, morphosyntactical structures across various language combinations (Monti et al., 2005), lexical density and variety (Sandrelli & Bendazzoli, 2005), and linguistic tendencies and patterns related to gender (Russo, 2018). EPICG, complied in the EXMARaLDA format (Schmidt & Wörner, 2009), aligns its audio signals or audio track with discourse annotation and interpretations. Research based on the multimodal components within EPICG sheds light on the hypothesised effects of short EVS (Defrancq, 2015). DIRSI, which includes annotations for mispronounced, truncated words, and units of meaning, provides an avenue to harness multimodality. This enables the consideration of speaking time and the rate of delivery, offering insights into participants' communicative power and their ability to express themselves effectively at international conferences (Bendazzoli, 2017). EPTIC has reintroduced punctuation marks to accurately capture speakers' intonation patterns. Additionally, other corpus-based studies rooted in the intricate relationship between various modalities within data from the European Parliament permit investigations into the impact of informational load on disfluencies (Plevoets and Defrancq, 2016b).

In the context of Chinese/English conference interpreting corpus, Chinese researchers have curated substantial and authentic English/Chinese corpora. These corpora consist of

speeches delivered by prominent figures from leading press outlets, along with their corresponding interpretations, which serve as valuable resources for qualitative and quantitative analysis (e.g., Hu and Tao, 2013; Pan, 2019). Among these resources, the Chinese English Political Interpreting Corpus (CEPIC) stands out, encompassing approximately 6.5 million words and enriched with POS tagging and various prosodic and paralinguistic features (Pan, 2019). Researchers have leveraged this corpus to explore various topics such as interpreting strategies and norms (B. Wang, 2012), normalization and explicitation (Hu & Tao, 2013), modal patterns in interpreted and translated discourses (Fu, 2016), as well as language specificity (B. Wang & Zou, 2018).

While some of these corpora are publicly accessible to the entire community for systematic analysis, others are accessible exclusively to researchers who are actively involved in developing multimodal conference interpreting corpora for specific academic purposes. In recent years, there has been a growing trend toward employing a multimodal approach in corpus-based research. On the paralinguistic level, researchers have explored the oral aspects of interpreting discourse (Han et al., 2020; Mead, 2000; B. Wang & Li, 2015; Yang, 2018), aligning with research on fluency in second language learning. On the non-verbal level, recent studies have investigated the positive influence of non-verbal paralanguage on meaning transfer in consecutive interpreting, involving the annotation of kinesic information such as facial expressions and gestures (Ouyang, 2020). Some studies have specifically focused on eye movements and gestures, shedding light on embodied cognition in simultaneous and consecutive interpreting (Stachowiak-Szymczak, 2019). In the domain of dialogue interpreting, there is increased attention to kinesics and proxemics, encompassing gesture, gaze, manners, and postures. For example, Tiselius and Sneed (2020) explore gaze patterns in dialogue interpreting, considering interpreters' actions and translation direction. Gao & Wang (2017) propose a multi-layer analytical framework comprising four categories of semiotic resources (written transcripts of utterances, auditory properties, video semiotics, and context) for the study of dialogue interpreting.

2.3 Methods and tools for transcription and data annotation

To investigate most interpreting phenomena, a fundamental requirement is the transcription of source speeches and interpretations, along with the annotation of prosodic or kinesic properties. The size of the corpus, which raises questions about representativeness and complexity in the transcription and annotation of acoustic signals, presents an ongoing challenge (Falbo, 2018). Therefore, researchers must strike a balance between the accuracy of the source and the adequacy of meeting the needs of corpus users (Bernardini et al., 2018). With the exception of EPIC, which underwent automatic transcription using speech recognition software followed by manual cross-checking, most corpora were orthographically transcribed according to standard transcription conventions. Depending on users' familiarity with different software tools, transcription can be conducted using such tools as Partitur Editor in EXMARaLDA (Schmidt and Wörner, 2009), Praat (Boersma & Van Heuven, 2001), or Transcriber (Barras et al., 2001). Audio or video tracks are imported into the transcription tool, allowing users to enter, edit, and produce time-aligned transcriptions linked to digital recordings. Additionally, transcription can be automated using EXMARaLDA via WebLicht as a Service to access the "speech-to-text" service API key.

Linguistic annotation involving POS tagging and lemmatization is accomplished using

fully automatic software programs such as Treetagger (Schmid, 1999) and CLAWS (Garside, 1987), or the POS Tagging module in the NLTK package in Python, which supports multiple languages. However, when it comes to the extraction of paralinguistic information in interpreting studies, there is currently no available tool for the automatic annotation and analysis of prosodic features. In the field of interpreting studies, researchers often resort to tools designed for speech analysis or building spoken language corpora to accommodate multimodality. For instance, the multi-layer corpus generated by EXMARaLDA allows for the simultaneous presentation of different tiers for each speaker, along with the assignment of absolute time values to corresponding recordings and the annotation of utterances accompanied by gestures or facial expressions in both simultaneous and dialogue interpreting. Multi-layer corpus can be utilized for manual measurement of Ear-Voice-Span (EVS). A common practice for identifying and annotating disfluency involves the conversion of acoustic signals into a visualised wave pattern, using software tools such as PRAAT or Cool Edit Pro. However, this process is time-consuming and labour-intensive.

To build a parallel corpus for conference interpreting, the alignment of the source and target transcripts at the level of sentences is a formidable task. It is acknowledged that sentences are inadequate for segmenting spoken language (Pietrandrea et al., 2014). As a result, "one-to-one correspondence between source and target segments is often missing in such corpora as EPIC and EPICG" (Bernardini et al., 2018, 29). The commonly used application to perform automatic alignment, such as Hunalign (Varga et al., 2007) built in Intertext Editor (Vondřička, 2014), can generate bilingual sentence pairs automatically aligned according to their sentence sequences.

2.4 Applicable technologies from Natural Language Processing

In today's technological era, where automation plays an increasing role across diverse sectors, spoken language and interpreting remain domains where full automation has not been achieved. Specifically, capturing the nuances and intricacies at the paraverbal level within multimodality presents a significant challenge. While no software can claim complete accuracy in extracting multimodal information in interpreting, certain technological advances, notably Automatic Speech Recognition (ASR), have made significant strides.

Automatic Speech Recognition (ASR) has its roots in the intertwined disciplines of computer science and linguistics. Its primary function is to convert spoken content into textual format. At its core, ASR relies heavily on machine learning, with particular emphasis on deep learning methodologies such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs). The advancements in these areas have enabled ASR to achieve substantial precision.

A variety of leading companies offer Automatic Speech Recognition (ASR) solutions, each with its unique strengths and applications. Google Cloud's ASR is renowned for its exceptional accuracy, robustness, and extensive language and dialect support, making it a top choice for real-time transcription and natural language processing applications. Amazon Transcribe stands out with customizable language models, integration with the AWS ecosystem, and a strong reputation for producing high-quality transcriptions. Microsoft Azure's Speech Service impresses with its combination of accuracy and advanced features such as speaker diarization and voice biometrics, making it an excellent choice for transcription, virtual assistants, and telecommunication applications. IBM Watson's Speech to Text service offers

robust customization options, enabling businesses to fine-tune the ASR system to their specific needs, making it ideal for applications in healthcare documentation and customer service. OpenAI's Whisper ASR, built on deep learning techniques, is known for its versatility and high accuracy across multiple languages and accents, making it a reliable choice for transcription services and voice-controlled applications. Baidu's DeepSpeech stands out for its remarkable performance in Mandarin Chinese and its open-source nature, which has made it a go-to option for Mandarin transcription and other languages with success. In a benchmark study published by (Thormundsson, 2021) which evaluated several speech-to-text service companies, Google emerged as the most reliable option for transcription. It achieved an impressive accuracy rate of 84.46%. This accuracy rate is calculated based on the Word Error Rate (WER) of 15.54%, representing the percentage of errors for every 100 words transcribed. In comparison, Amazon Web Service achieved an accuracy rate of 83.12%, while Microsoft slightly lagged behind with an accuracy rate of 81.01%.

For researchers and developers, NVIDIA NeMo ASR, part of the NeMo toolkit, provides pre-trained models and extensive customization capabilities, making it a valuable choice for those involved in the exploration and development of speech recognition and synthesis applications. When considering these ASR solutions, factors like accuracy, language support, customization options, real-time capabilities, integration with other services, and pricing are crucial considerations, as the technology continues to evolve rapidly, with companies regularly introducing new features and improvements to meet the growing demand for advanced speech recognition solutions in various domains.

Constructing multimodal interpreting corpora presents another challenge: aligning the transcribed source text with the target text. The intricate challenge of text alignment across diverse languages has long captivated scholars and practitioners, especially due to its significance in machine translation and computational linguistics. In interpreting studies, alignment not only paves the way for the creation of parallel corpora but also underpins the methodology for this project's automated assessment of interpreting quality. NLP has evolved to facilitate the alignment task which extends beyond mere syntactic alignment; it probes into recognizing and capturing the deeper essence, intricacies, and the semantic shades of every sentence.

LASER (Artetxe & Schwenk, 2019) emerges as a significant tool in this domain. Its usability lies in mapping multilingual texts onto a singular vector space. This design allows for the identification and positioning of sentences that convey similar semantic undertones, irrespective of their linguistic origin. By championing a communal semantic architecture, LASER transcends the limitations of individual languages, thereby introducing a methodology that holds universal applicability. Such an approach ensures that whether the linguistic pair in question is English/Chinese or another combination, LASER provides a methodology to measure semantic alignment. Vecalign (Thompson & Koehn, 2020) carves its niche as a superior bilingual sentence alignment tool. By leveraging the similarities between sentence embeddings, Vecalign achieves precise bilingual alignments. Vecalign can align sentences from parallel documents based on their semantic equivalence. The tool operates with the understanding that content in parallel narratives, even when articulated in different languages, will mirror similar thematic sequences.

Bleualign stands out by integrating the BLEU metric, commonly employed for machine

translation evaluations. It refines alignments by maximizing the BLEU score (Tiedemann, 2011). In parallel, Hunalign, another venerable tool, ingeniously merges dictionary-driven techniques with sentence length analysis to generate alignments (Varga et al., 2007). Each tool brings a distinct contribution to alignment. Gargantua hinges on dynamic programming (Braune & Fraser, 2010) for its precision, while ParAlign incorporates paraphrase-based metrics (Rognes, 2001). On the other hand, FastAlign, celebrated for its rapidity, relies on the principles of phrase-based machine translation (Chahuneau et al., 2013). Innovations like YASA are designed to address significant disparities in parallel content (Tiedemann, 2011) Meanwhile, MUSE and mBERT further expand the alignment toolkit. MUSE focuses on curating cross-lingual embeddings under supervised and unsupervised settings (Conneau et al., 2018), whereas mBERT, a multilingual iteration of BERT, facilitates semantic alignments (Devlin et al., 2019).

2.5 Research questions

Despite the diverse modes and settings explored in previous work of interpreting corpus construction, there are common challenges in the methods used for transcription, prosodic annotation, and alignment. Specifically, the time-consuming nature of ethnographic transcription can limit the quantity of data collection, especially when a single researcher is responsible for building the corpus. Based on our literature review, particularly the construction of EPIC (Russo et al., 2012), as well as the availability of transcription tools and recent advancements in automatic speech recognition technology, it appears that ASR holds potential for application in constructing interpreting corpora. However, questions remain about its feasibility, which leads us to our first set of research questions.

Furthermore, the repetitive process of tagging on the paralinguistic level affects the ability to investigate multimodal phenomena beyond verbal expression. Drawing inspiration from Poyatos' (2002) concept of sign-conveying verbal and nonverbal systems, this study also aims to explore methods for automatically extracting paralinguistic information.

The available methods for automatic alignment can be categorized into types such as N-gram, phrase-based metrics, dictionary-driven techniques, and sentence-level embeddings, as indicated by our literature review. What sets interpreting apart from translation is that, in interpreting, the overall message is conveyed with primary and secondary information retained, in contrast to a strict word-for-word or phrase-for-phrase alignment. Consequently, this study seeks appropriate solutions based on automatic text alignment technology to expedite the construction of multimodal interpreting corpora. The study aims to offer preliminary insights into the following sets of research questions (RQ):

- RQ 1: Can automated speech recognition technology (ASR) be effectively applied for transcribing conference interpreting? What is the error rate in using typical ASR for transcribing interpreting renditions?
- RQ 2: Is it possible to automatically identify or extract certain paralinguistic features of interpreting more efficiently than by visually analysing the wave patterns in oscillograms for conference interpreting? If yes, how can the features be extracted?
- RQ 3: What methods can be employed to align source and target transcripts automatically at the sentence level for the construction of interpreting corpus?

These three sets of questions revolve around fundamental steps in corpus construction. The study aims to explore how natural language processing can be structured in various ways

to facilitate problem refinement and the encoding of necessary tasks in the corpus-building process. In essence, the answers to the above research questions offer potential solutions to automatic transcription, paralinguistic information identification, and automatic alignment. They shed light on the realisation of multimodal corpora by introducing new methods for extracting and annotating paralinguistic features, which warrant additional empirical description and characterisation for a comprehensive understanding.

3. Feasibility of automatic transcription for building interpreting corpora

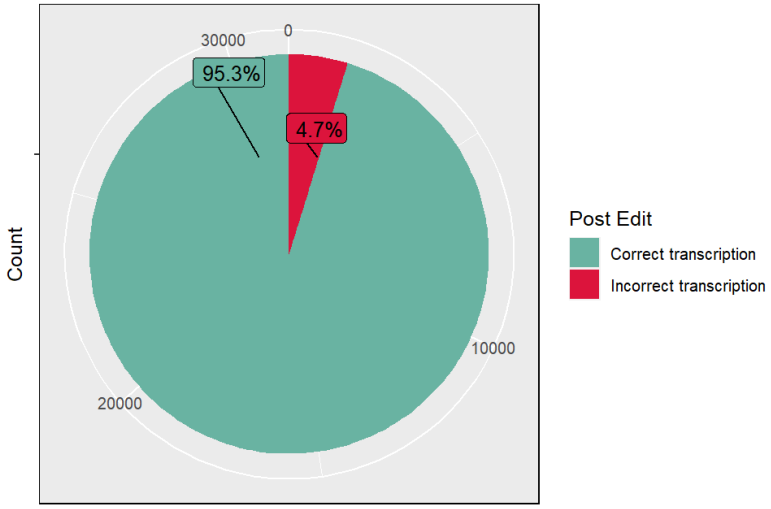
As neural networks have demonstrated significant advancements in speech recognition tasks, researchers have begun to explore the potential of ASR to address some of the challenges in corpus linguistics. This has the potential to provide support to various stakeholders, including interpreting researchers, educators, and practitioners, in their endeavours to investigate language using large, machine-readable corpora. In this section, we explore several key questions to assess the feasibility of employing automatic transcription in the construction of interpreting corpora: 1) What is the error rate in using typical ASR for transcribing interpreting renditions? 2) Besides generating text transcripts, what additional information can this technology extract, and how can it be leveraged for corpus construction? 3) Among the prevalent software options, which ASR service should be employed when transcribing interpretations across different modes and settings? 4) What are the limitations of using this technology beyond issues related to transcription accuracy?

The data used in this study are recorded interpreting performances of graduating trainees from a postgraduate professional interpreting programme. These trainees, all majoring in interpreting, are prospective professionals who have undertaken two semesters of training in consecutive and simultaneous interpreting. Their first language (L1) is Mandarin Chinese, and their second language (L2) is English. The dataset consisted of forty-nine recordings of English – Chinese consecutive interpreting. These interpreting performance recordings were collected at three key assessment points throughout the year-long postgraduate interpreting programme: at the midpoint and at the end of Semester 1, and at the end of Semester 2. During Semester 1, participants consecutively interpreted a speech lasting approximately 4.5 minutes at the mid-term assessment, and a speech lasting approximately 5.5 minutes in the final assessment. At the end of Semester 2, they consecutively interpreted a more specialised and information-dense speech of about seven minutes in length. All speeches were delivered spontaneously without scripts. All assessments were conducted in an interpreting training lab replicating real-world conditions for consecutive interpreting.

This study utilizes Application Programming Interfaces (APIs) for the speech-to-text services provided by IBM and Google, which have been verified to offer high transcription accuracy with low Word Error Rate (WER). The use of these APIs is aimed at optimizing the process for the automatic construction of interpreting corpora. The findings reveal that human post-editing, aimed at correcting inaccurate transcriptions constitute 4.7% of the entire transcripts, implying an accuracy rate of 95.3% for automatic transcription by IBM (see Figure 1). Google's ASR is the only tool capable of recognising accent variations in spoken Chinese, encompassing Mandarin, Cantonese, and Taiwanese, as well as idiosyncratic language and grammar differences, even within the same language.

Figure 1

Accuracy of automatic transcription



Regarding the results of using "Speech-to-text" cloud service provided by IBM and Google via APIs, the transcription includes time offset values (timestamp) and Confidence Measure (CM) in JavaScript Object Notation format (JSON), as requested (refer to Figure 2). Time offset values indicate the beginning and ending time of each spoken word recognised in the audio. CM serves as a performance index, representing a score that evaluates the reliability of recognition results. A higher CM score denotes greater confidence in the accuracy of the transcription. In contrast to time-aligned format in EXMARaLDA, the time offset values in the JOSN file can be extracted in batches, allowing for calculation of silence and articulation time. This enables the study of how the interpreter modulates speech by evaluating the intervals between timestamps. Additionally, CM can be employed to assess potentially "misheard" words by the machine, as low CM scores may indicate sections where the speech was unclear or where the interpreter may have mumbled, hesitated, or deviated from standard language structures. Therefore, it is essential to explore whether low CM scores can function as an indicator of unintelligibility in the ASR system's interpretation of the sound signal.

Figure 2

Example of transcription results with timestamp and CM in JSON format

```
      "transcript": "我 个人 是一个 女权主义 者 我 也 认为 作 作 一个 女权主义 主义者 是 一件 很重要 的 事情 你们 都 知道 在 过去 的 一周 里 我们 过了 万圣节
    },
    {
      "transcript": "我 个人 是一个 女权主义 者 我 也 认为 做 做 一个 女权主义 主义者 是 一件 很重要 的 事情 你们 都 知道 在 过去 的 一周 里 我们 过了 万圣节
    }
   ]
  },
  {
   "final": true,
   "alternatives": [
    {
     "transcript": "还 庆祝 着 平等 公平 ",
     "confidence": 0.77,
     "timestamps": [
      [
       "还",
       24.11,
       24.38
      ],
      [
       "庆祝",
       24.41,
       24.82
      ],
      [
       "着",
       24.82,
       25.13
      ],
      [
       "平等",
       25.34,
       25.93
      ],
      [
       "公平",
       26.11,
```


As previously mentioned, Google's technology is the preferred choice when a text-form transcript is required due to its relatively high accuracy. Additionally, Google's ASR is capable of transcribing simultaneous interpreting, as it can recognise multiple speakers in the same audio clip via API service. In terms of time offset values, IBM returns more accurate timestamps that preserve filled pauses in interpreted texts, whereas Google combines unfilled pauses with the duration of spoken words. Consequently, the timestamps provided by IBM are more suitable for further research based on a prosodically annotated corpus.

While technology has made significant progress in generating preliminary drafts that could potentially replace orthographical transcription, this methodology falls short when it comes to automatically inserting punctuation marks in the correct positions within sentences, particularly in interpreted texts during simultaneous interpreting. This limitation arises because ASR identifies periods during relatively long pauses between words, assuming that speakers use oral punctuation correctly to allow the audience to reflect on their speech. However, silent pauses occur when interpreters hesitate for various reasons such as lexical or syntactical planning, resulting in unwanted automatic segmentation in the interpreted texts and additional work required for alignment.

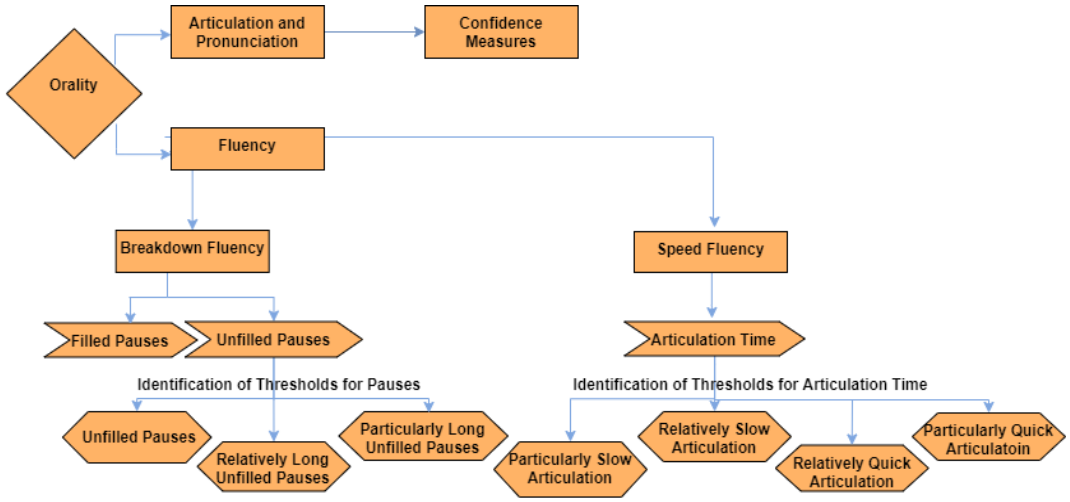
4. Development and implementation of an automatic annotation scheme on the paralinguistic level

At a paralinguistic level, the transcription of speech production focuses on delivery-related indicators. Prosodic properties associated with interpreting delivery can be automatically generated through rule-based programming once the relevant features have been identified through statistical analysis. In this study, the identification of paralinguistic information for the multimodal corpus is based on a multi-layer model (Figure 3) inspired by Poyatos (2002), which includes elements from his acoustic and visual sign-conveying system, as well as insights from fluency research by (Tavakoli & Skehan, 2005). Building on the existing literature that has brought to light the importance of delivery, such as Bühler (1986), Collados Aís, (2002); Kurz (1993), Kurz and Pöchhacker (1995), Moser (1996), Tavakoli and

Skehan (2005), Pöchhacker and Zwischenberger (2010), delivery features are related to fluency, as well as articulation and pronunciation. Articulation and pronunciation refer to the ability to interpret with intelligible and correct sounds of words that the audience can easily understand. Empirical evidence substantiates the use of Confidence Measure (CM) as an index for assessing articulation and pronunciation, where low CM values indicate mispronunciation (refer to the findings in section 4.2).

Figure 3

A multi-layer model for automatic construction of a multimodal corpus (X. Wang & B. Wang, 2022)



In terms of fluency, recent empirical research findings (Yang, 2018; Han et al., 2020) align with the homogeneous categorisation of fluency into breakdown, speed, and repair fluency as introduced by Tavakoli and Skehan (2005). Breakdown fluency assesses the extent of interruptions in speech caused by pauses and filled pauses. In this research study, we employ a rigorous statistical analysis to objectively quantify temporal measures that characterize breakdowns in fluency. These measures encompass unfilled pauses (UP), relatively long unfilled pauses (RLUP), particularly long unfilled pauses (PLRP), and filled pauses. The precise temporal definitions of these terms can be found in Table 1.

The second subparameter of utterance fluency, known as speed fluency, quantifies the number of words that a speaker can accurately articulate per minute. However, calculating the speech tempo for each sentence is hindered by the previously mentioned unreliable punctuation-delimited clauses. Conversely, it is relatively simple to detect and illustrate variations in tempo by timestamping each word as it is spoken. Consequently, articulation time per character or word is calculated for both renditions to measure speed fluency. Specifically, this subparameter is quantified using four temporal variables: particularly slow articulation (PSA), relatively slow articulation (RSA), relatively quick articulation (RQA), and particularly quick articulation (PQA). Please refer to Table 1 for the precise temporal definitions of these terms.

Repair fluency pertains to the occurrence of corrections and repetitions within a speech. As defined by Tavakoli and Skehan (2005), repair fluency encompasses reformulation, replacement, false starts, and repetition. Nevertheless, transcription of repair fluency cannot be

achieved using an automated tool or rule-based programming, as current natural language processing technology is unable to generate these features. Accommodating shifts in a speaker's thoughts remains a challenging task for artificial intelligence.

4.1 Identification of (dis)fluency based on a statistical calculation

In this study, unfilled pauses are categorized as breakdown fluency, while articulation time is classified as speed fluency based on their duration. All data concerning silence and articulation are computed using a Python script that makes use of time offset values. This script is available on GitHub for the automatic extraction of disfluency for other future studies. By identifying a threshold through statistical analysis of the five-number summary of interpreting data, this study identifies outliers that exhibit significant deviation from the rest of the silence data as RLUP and PLUP and those related to articulation time as RSA, PSA, RQA, and PQA. Table 1 provides a summary of the defined fluency measures.

Table 1

List of temporal measures of fluency and their brief definition

Fluency Parameters	Definition and Calculation
Unfilled pauses (UP)	Unfilled pauses equal to and longer than 250 milliseconds and shorted than but not outliers
Relatively long unfilled pauses (RLUP)	Unfilled pauses longer than Q3 +1.5 * IQR and shorter than and equal to Q3 + 3 * IQR
Particularly long unfilled pauses (PLUP)	Unfilled pauses longer than Q3 + 3 * IQR
Relatively slow articulation (RSA)	Articulation time per syllable longer than Q3 +1.5 * IQR and shorter than and equal to Q3 + 3 * IQR
Particularly slow articulation (PSA)	Articulation time per syllable longer than Q3 + 3 * IQR
Relatively quick articulation (RQA)	Articulation time per syllable shorter than Q1 -1.5 * IQR and longer than and equal to Q1 - 3 * IQR
Particularly quick articulation (PQA)	Articulation time per syllable shorter than Q1 - 3 * IQR

Note. Q1: the first quartile; Q3: the third quartile; IQR: the interquartile range (Q3 – Q1)

In all corpora, filled pauses are identified as '啊(uh)', '嗯(mm)', and '呃(er)' in the target-language output in the English-Chinese Parallel Corpus of Consecutive Interpreting and Parallel Corpus of Simultaneous Interpreting. In the Chinese-English direction, filled pauses refer to 'uh', 'uhm', 'euh' and 'euhm'. Extraction of filled pauses is quick and easy thanks to ASR, as they have been tagged as '%HESITATION%' in the transcription results, regardless of form in which they uttered in the audio. The study assigns these mark-ups by executing commands in Python scripts.

The automated extraction of (dis)fluency information, using newly proposed parameters, opens up two avenues of research. Firstly, researchers can explore the cognitive aspects of interpreting tasks. Studies have investigated whether the complexity of interpreting tasks is reflected in established utterance measures and patterns of disfluency in speech. Recent years have seen a growing body of empirical research aiming to model the relationship between temporal characteristics and cognitive load. Secondly, these newly defined measures are useful for pedagogical purposes. Researchers can investigate whether these measures are effective in assessing fluency in interpreting studies. If an optimal set of acoustic measures can accurately

predict human-judged fluency, it can enable the automation of score prediction and provide instant feedback on disfluency information. This can be valuable for formative and diagnostic assessment purposes.

4.2 Predictability of confidence measure for mispronounced words

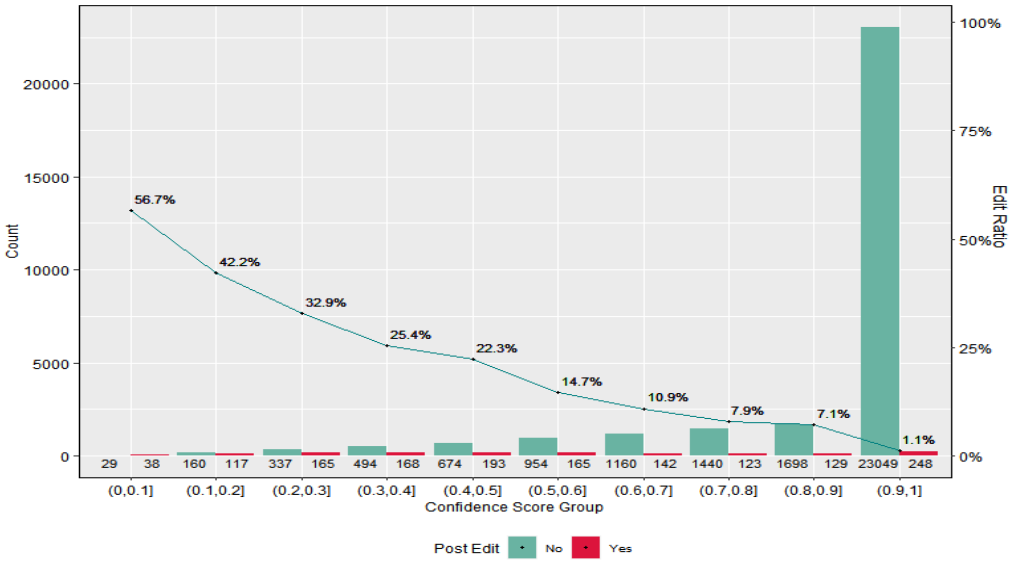
In this study, we investigate whether confidence measures (CM) obtained through the Speech-to-text Service API can be used to identify instances of unintelligibility in interpretations. CM is a score ranging from zero to one, assigned to each word and individual sentence. It serves as an indicator of the likelihood that the spoken words are accurately recognized. A higher CM score corresponds to a higher level of reliability in the transcription results. CM is generated by utilizing a combination of trained predictor features related to acoustics, syntax, and semantics, which are collected during the decoding process (Huang et al., 2013; Jiang, 2005).

Two research questions need to be addressed. First, what is the predictability of CM as an indicator for mispronounced words? Second, if CM can indeed predict mispronunciations, what threshold value should be used to identify unintelligibility? The experimental results demonstrate that CM is a reliable indicator, achieving an accuracy of 95.3% in identifying mispronounced words. Additionally, a CM value of 0.321 has been identified as the optimal threshold for detecting unintelligibility.

The study comprises forty-nine consecutive interpreting renditions performed by professional trainees. This dataset was previously used to test the viability of automatic transcription for building interpreting corpora, albeit with different annotations in this case. Unintelligibility, referring to words that were pronounced incorrectly or unclearly, was manually identified and labelled as binary data (clear/unclear errors) by annotators. The collected quantitative data were then analysed using a binary logistic regression model with CM as the sole independent variable. To assess the model's performance and mitigate the risk of overfitting or selection bias, the study employed Receiver Operating Characteristic (ROC) curves and K-fold cross-validation. Furthermore, the study addressed data imbalance by downsampling the majority category, as the data exhibited an imbalance between the two types, potentially affecting model performance.

Figure 4

Distribution of CM scores and labelling rate



The results indicate that the majority of words are accurately recognised, with only 4.23% (1,629 out of 38,444) annotated due to unclear or incorrect pronunciation, and 0.13% failing to be transcribed accurately for other reasons, such as homophones or illogical expressions. Figure 4 illustrates the distribution of confidence levels and labelling rates in ten equally divided groups between zero and one. The bar chart depicts the number of words correctly transcribed (green bars) and those labelled as unclear (red bars) within each interval. The number of words in the green bars steadily increases from the interval of (0, 0.1] to (0.8, 0.9], experiencing a significant surge within the last interval. Similarly, the red bar chart remains relatively stable in the first nine groups but exhibits a slight rise in the last group. Overall, there is a noticeable upward trend in the number of accurately recognised words in speeches, with 73.2% falling within the interval with the highest confidence scores. Labelling rates for each interval are provided in Figure 4. It is evident that labelling rate is inversely related to confidence level: as one quantity decreases, the other increases. This relationship suggests that the lower the confidence score, the higher the editing rate should be, reflecting lower reliability in the recognition result reflected by CM score for speech with less clarity.

Table 3

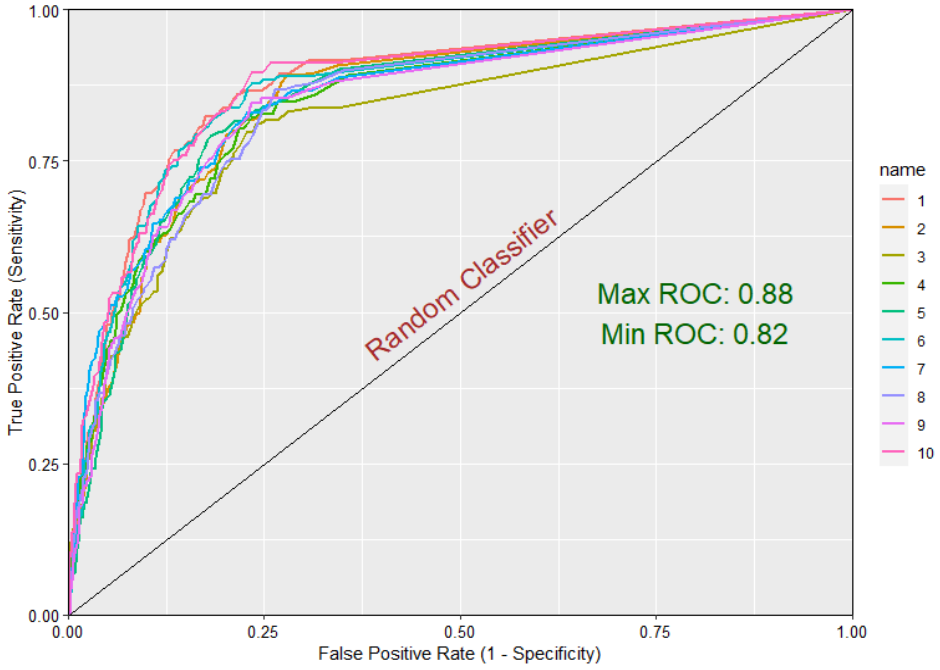
Cross-validation for logistic regression models

No	Cut Off	Precision	Sensitivity	Specificity	AUC	P-Value
1	0.433	0.182	0.824	0.825	0.876	9.0549e-146
2	0.243	0.157	0.891	0.720	0.853	2.9945e-147
3	0.287	0.143	0.811	0.760	0.820	1.3913e-151
4	0.319	0.156	0.822	0.767	0.847	3.8166e-151
5	0.424	0.191	0.790	0.818	0.851	4.7693e-146
6	0.317	0.162	0.877	0.766	0.866	1.1917e-143
7	0.331	0.146	0.830	0.772	0.855	2.7539e-147
8	0.254	0.142	0.867	0.737	0.845	6.5333e-150
9	0.308	0.141	0.846	0.767	0.845	3.7370e-151
10	0.292	0.136	0.895	0.766	0.877	2.2070e-149

No	Cut Off	Precision	Sensitivity	Specificity	AUC	P-Value
Average	0.321	0.155	0.845	0.770	0.854	

Figure 5

ROC curves of cross-validation with regression models



To further assess the predictability of unintelligibility with CM, this study constructs logistic regression models for evaluating classifier performance through cross-validation. The results depicted in Figure 5 illustrate how these curves correspond to different K-fold cross-validation datasets. ROC curves that are closer to the top-left corner indicate good performance, characterized by a false positive rate of zero and a true positive rate of one. Model 10 performs the best, achieving an Area Under the Curve (AUC) of 87.7%, with a specificity of 76.6% and sensitivity of 89.5%, while model 3 performs the least well, with an AUC of 82%, a specificity of 76%, and sensitivity of 81.1% (as shown in Table 3). On average, the AUC of the curves is 85.4%, with a specificity of 77% and sensitivity of 84.5%. The AUC results are statistically significant, indicating that 84.5% of unintelligible words have been correctly identified, while 23% of intelligible words have been misclassified as unintelligible. These AUC results are considered good, suggesting that the regression models exhibit ideal separability measures. The classifier outputs remain relatively stable even with variations in the training data. Thus, it can be concluded that a suitable threshold for defining unintelligibility is an average cut-off value of 0.321, where any word with a confidence measure below 0.321 should be labelled as unintelligible.

The exploration conducted above holds significance because statistical results support CM as an indicator of articulation and pronunciation in the field of interpreting studies. This research direction has the potential to stimulate the automated assessment of delivery in interpreting studies. Paraverbal information, which plays a vital role in assessing delivery, can be quantified automatically using Python scripts with the support of ASR results. Essentially, it becomes possible to develop a machine learning model that incorporates CM and fluency parameters for predicting human ratings of delivery. CM, when integrated into such a model,

contributes to the creation of a technically robust assessment system capable of producing valid results when combined with other relevant parameters. The prediction outcomes derived from this model can offer essential insights for educators and researchers, aiding in decision-making for both pedagogical and research purposes.

Furthermore, investigating the relationship between CM, pronunciation, and articulation has implications for the design and development of automated assessment systems for information fidelity of interpretation. The assessment of information fidelity processes relies heavily on the accuracy of speech recognition results, necessitating the input of precise transcriptions into an automatic scoring system. However, the presence of unintelligible words in the target-language output can lead to inaccurate transcription results. Therefore, to enable automated assessment with minimal human intervention, it is proposed that words with a CM value below the cut-off point (0.321) should be excluded. This ensures that the content conveyed in target renditions can be accurately compared to the source speech. Consequently, automated CM extraction is expected to facilitate researchers in gaining insights into one of the explicit quality criteria and provide an approach that can be readily implemented to assess pronunciation, articulation, and delivery.

5. Automatic alignment in constructing parallel interpreting corpora

This section outlines the process of automatically alignment for constructing parallel interpreting corpora based on cross-lingual semantic similarity between the transcribed source and target texts. The procedure involves two key steps: automatic text segmentation and automatic alignment, which utilize deep learning technology. As an example, we'll consider the "Speech-to-text" service provided by Google.

The initial step involves segmenting the transcribed source speech and interpretations automatically. This segmentation is achieved by identifying sentence boundaries and the presence of punctuation marks, such as periods or commas. This identification relies on detecting very long pauses during the automatic transcription process.

To facilitate parallel corpus mining, all sentence-level language pairs are jointly embedded in a shared space for representation. This process leverages LASER (Language-Agnostic SEntence Representations toolkit) (Artetxe & Schwenk, 2019) to generate embeddings (dense vector representation) for sentences in a way that similar sentences, regardless of the language they are in, have similar embeddings. This makes it possible to compare the contrast sentences between English and Chinese in the embedding space.

Next, Vecalign (Thompson and Koehn, 2020) tries to align sentences from the transcribed source and target text based on the distances between these points in the embedding space. The rationale is that similar sentences in different languages will have similar meanings, so their vectors will be close to each other in the embedding space. Vecalign employs an optimisation algorithm to make the alignments coherent and respect the order of the texts. For instance, if sentence A in the source text is aligned with sentence X in the target text, then sentence B (which comes after A) should be aligned with sentence Y that comes after X (or possibly still with X if B is a continuation of A).

To evaluate the viability and quality of automatic alignment, this study conducts a comparative analysis between the results automatic and manual alignment. The approach involves assessing three English-Chinese renditions by interpreting trainees at the sentence

level using both alignment methods. The goal is to determine the degree of similarity between the assessment outcomes obtained through these two alignment methods. High similarity between the assessment results suggests that automatic alignment closely aligns with manual alignment, indicating the feasibility of automatic alignment.

The three interpreting renditions were selected from a larger pool of forty-nine recordings, which were also used in the previous two experiments. It's important to note that these selections were not random; instead, they were chosen deliberately to represent a range of performance qualities within the forty-nine recordings. Specifically, the renditions chosen include the highest-rated, the lowest-rated, and a rendition with a median rating on the document level. This curated selection was made to ensure that the sentence-level data would follow a normal distribution. Additionally, the choice of three renditions helped conserve computational resources and reduced the amount of human labour required for sentence-level assessment. In the case of manually aligned language pairs, it is essential to segment and align both the source and target texts based on natural sentences before performing text alignment. As a result, there are a total of 206 sentence pairs from manual alignment and 156 from automatic alignment.

To assess the consistency of information fidelity in each language pair, two raters were recruited, both of whom are native Mandarin-Chinese interpreting trainers with fluency in English. These raters were selected based on their outstanding academic qualifications and extensive experience in interpreting, teaching, and assessing performances by interpreting trainees. The first rater holds a postgraduate degree in interpreting and currently serves as a consecutive interpreting lecturer at a prestigious university in China. The second rater, actively pursuing a PhD, is an experienced interpreter with more than four years of experience as an interpreting trainer at another Chinese university. The ratings provided by these two individuals showed a high level of agreement in assessing language pairs for both manual alignment (Cohen's Kappa = 0.87) and auto-alignment (Cohen's Kappa = 0.88).

One-way ANOVAs were conducted to examine whether there were significant differences between the results obtained from the automatic alignment and manual alignment methods. The analysis revealed that there were no statistically significant differences between the parallel sentences aligned using the two different methods ($F(1, 152) = 1.562$, $p = 0.213$). These statistical findings indicate the successful nature of the automatic alignment approach for aligning parallel English/Chinese sentences through hidden internal representations, even if the aligner is not 100% accurate. It is anticipated to become an essential component of computer-assisted translation tools. Furthermore, it can be utilized in the creation of parallel corpora for research in translation and interpreting studies, facilitating a wide range of analyses between source and target texts.

6. Conclusion

This study has developed new approaches to automatic construction of multimodal interpreting corpora, incorporating new parameters identified through new technologies. The process commences with the exploration of neural network technology and its application in automatic transcription, which yields written utterances and machine-readable values such as timestamp and confidence measure. These data can then be processed to generate multimodal information. The multi-layer model proposed in this paper draws inspiration from the concepts

of multimodality in information processing within interpreting studies, and of fluency in spoken language processing research and language learning investigations.

The corpora created through this methodology encompass two unidirectional parallel corpora: 1) "The Multimodal Corpus of English-Chinese Conference Interpreter Training" (MCECCIT), featuring 49 recordings by 24 trainees, and 2) "The Multimodal Corpus of Chinese-English Conference Interpreter Training" (MCCECIT), which includes 54 recordings performed by 39 trainees (see Appendix B).

In response to Research Question 1, the error rate when utilizing IBM's ASR service for transcribing English source speeches and their interpretation is found to be as low as 4.7%. While this necessitates some manual correction, it proves to be considerably more efficient compared to manual transcription methods. Regarding Research Question 2, prosodic properties, including filled pauses, unfilled pauses, and articulation rate, can be identified through statistical analysis and extracted automatically using our rule-based programming. Additionally, mispronounced words, a paralinguistic aspect, can be detected based on Confidence Measure with values lower than the average cut-off of 0.321. Regarding Research Question 3, the automated alignment of English/Chinese language pairs involves a sequence of steps: transcribed source speeches and interpretations are initially segmented into sentences using ASR, subsequently converted into embeddings with LASER, and finally aligned coherently with Vecalign to reflect the order of the source text. All detailed methods, implemented in Python scripts, are openly accessible on the author's GitHub repository: https://github.com/renawang26/Automatic-methods-for-Construction-of-Multimodal-Interpreting-Corpora.

The study addresses elements deemed important by many interpreting researchers for capturing paralinguistic information. This effort provides a method, rooted in corpus linguistics, to explore orality within discourse studies and offer insights into interpreting quality. Parameters related to oral traits such as disfluency, pronunciation, and articulation have been identified. The approach in this study is primarily an effective step towards reducing manual labour in building multimodal interpreting corpora by applying automated methods.

Despite the implications, the study has several limitations. Although our study has showcased the promising performance of individual technological components for automatic construction of interpreting corpora, when these components are combined into a singular processing pipeline, the efficiency and accuracy might be compromised. The integrated system might not perform as robustly as the individual component did. Consequently, human intervention, including corrections and post-editing, remains a crucial step in ensuring the successful construction of a reliable corpus.

Additionally, as the technologies and methods employed in this study have been specifically tailored for the English/Chinese language pair, the results obtained in this particular context might not be directly transferable to other language pairs. For other low-resource languages, which have fewer technological tools and resources available, our findings might not be indicative of the potential outcomes. Therefore, researchers and practitioners should exercise caution when extrapolating our results to other linguistic pairs.

In addition to transcribing acoustic information, a more significant challenge lies in the automatic identification and annotation of kinesics, including gaze, facial expressions, gestures, and posture, by incorporating emerging multi-camera system technology. This issue warrants

further exploration to enhance the multi-layer model for constructing a multimodal corpus. Another minor issue also arises from the incorrect punctuation assigned by speech recognition to transcripts. For example, it is impossible to investigate whether an extended average sentence length contributes to processing difficulties, as ASR lacks the capability to accurately determine sentence boundaries based on acoustic and lexical evidence.

References

Artetxe, M., & Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610. https://doi.org/10.1162/tacl_a_00288

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1–2), 5–22.

Bendazzoli, C. (2012). From international conferences to machine-readable corpora and back: An ethnographic approach to simultaneous interpreter-mediated communicative events. In *Breaking Ground in Corpus-Based Interpreting Studies* (Vol. 147, pp. 91–117). Springer.

Bendazzoli, C. (2017). Benefits and drawbacks of English as a lingua franca and as a working language: The case of conferences mediated by simultaneous interpreters. In *English in Italy. LInguistic, Educational and Professional Challenges* (pp. 119–141). FrancoAngeli.

Bendazzoli, C. (2018). Corpus-based interpreting studies: Past, present and future developments of a (wired) cottage industry. In *Making way in corpus-based interpreting studies* (pp. 1–19). Springer.

Bendazzoli, C., Bertozzi, M., & Russo, M. (2020). From text to multimodal resources: Advancing interpretation research from an already existing corpus. *META*, *65*(1), 210–235.

Bendazzoli, C., Russo, M., & Defrancq, B. (2018). Corpus-based Interpreting Studies: A booming research field. *INTRALINEA ON LINE TRANSLATION JOURNAL*, *20*, 1–2.

Bernardini, S., Ferraresi, A., Russo, M., Collard, C., & Defrancq, B. (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In *Making way in corpus-based interpreting studies* (pp. 21–42). Springer.

Biagini, M. (2012). Data collection in the courtroom: Challenges and perspectives for the researcher. In *Breaking ground in corpus-based interpreting studies* (Vol. 147, pp. 231–252). Springer.

Boersma, P., & Van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341–347.

Braune, F., & Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. *Coling 2010: Posters*, 81–89.

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. *Multilingua*, *5*(4), 231–235.

Bührig, K., Kliche, O., Meyer, B., & Pawlack, B. (2012). The corpus "Interpreting in Hospitals": Possible applications for research and communication training. In *Multilingual corpora and multilingual corpus analysis* (pp. 305–315). John Benjamins.

Chahuneau, V., Smith, N. A., & Dyer, C. (2013). Knowledge-rich morphological priors for bayesian language models. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1206–1215.

Collados Aís, Á. (2002). Quality assessment in simultaneous interpreting: The importance of nonverbal communication. *The Interpreting Studies Reader*, 327–336.

Collados Aís, Á., Fernández Sánchez, M. M., Iglesias Fernández, E., Pérez-Luzardo, J., Pradas Macías, E. M., Stévaux, E., Blasco Mayor, M. J., & Jiménez Ivars, A. (2004). Presentación de Proyecto de Investigación sobre Evaluación de la Calidad en Interpretación Simultánea (Bff2002-00579). *Actas Del IX Seminario Hispano-Ruso de Traducción e Interpretación. Moscú (Moscow): Universidad Estatal Lingüística de Moscú (MGLU)*, 3–15.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). *Word Translation Without Parallel Data* (arXiv:1710.04087). arXiv. https://doi.org/10.48550/arXiv.1710.04087

Dal Fovo, E. (2018). The use of dialogue interpreting corpora in healthcare interpreter training: Taking stock. *The Interpreters' Newsletter*, *23*, 83–113.

Defrancq, B. (2015). Corpus-based research into the presumed effects of short EVS. *Interpreting*, *17*(1), 26–45. https://doi.org/10.1075/intp.17.1.02def

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Falbo, C. (2012). CorIT (Italian Television Interpreting Corpus): Classification criteria. In S. S. Francesco & C. Falbo (Eds.), *Breaking Ground in Corpus-Based Interpreting Studies* (pp. 155–186). Springer. https://doi.org/10.3726/978-3-0351-0377-9/6

Falbo, C. (2018). La collecte de corpus d'interprétation: Un défi permanent. *Meta : journal des traducteurs / Meta: Translators' Journal*, *63*(3), 649–664. https://doi.org/10.7202/1060167ar

Fu, R. (2016). Comparing modal patterns in Chinese-English interpreted and translated discourses in diplomatic setting: A systemic functional approach. *Babel*, *62*(1), 104–121. https://doi.org/10.1075/babel.62.1.06fu

Gao, F., & Wang, B. (2017). A multimodal corpus approach to dialogue interpreting studies in the Chinese context: Towards a multi-layer analytic framework. *The Interpreters' Newsletter*, *22*, 17–38. https://doi.org/10.13137/2421-714X/20736

Garside, R. (1987). *The CLAWS word-tagging system. In: R. Garside, G. Leech & G. Sampson (eds.), The Computational Analysis of English: A corpus-based approach. London: Longman*.

Gavioli, L. (2015). On the distribution of responsibilities in treating critical issues in interpreter-mediated medical consultations: The case of "le spieghi (amo)". *Journal of Pragmatics*, *76*, 169–180.

Halliday, M. A. K. (2014). Language as social semiotic. *The Discourse Studies Reader. Amsterdam: John Benjamins*, 263–272.

Han, C., Chen, S., Fu, R., & Fan, Q. (2020). Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. *Interpreting.*

International Journal of Research and Practice in Interpreting, *22*(2), 211–237. https://doi.org/10.1075/intp.00040.han

House, J., Meyer, B., & Schmidt, T. (2012). CoSi-A corpus of consecutive and simultaneous interpreting. *Multilingual Corpora and Multilingual Corpus Analysis*, 295–304.

Hu, K., & Tao, Q. (2013). The Chinese-English Conference Interpreting Corpus: Uses and Limitations. *Meta: Journal Des Traducteurs / Meta: Translators' Journal*, *58*(3), 626–642. https://doi.org/10.7202/1025055ar

Huang, P.-S., Kumar, K., Liu, C., Gong, Y., & Deng, L. (2013). Predicting speech recognition confidence using deep learning with word identity and score features. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7413–7417. https://doi.org/10.1109/ICASSP.2013.6639103

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech Communication*, *45*(4), 455–470.

Kress, G. (2009). *Multimodality: A social semiotic approach to contemporary communication*. Routledge.

Kress, G., & Van Leeuwen, T. (2001). *Multimodal discourse, The Modes and Media of Contemporary Communication*.

Kurz, I. (1993). Conference interpretation: Expectations of different user groups. *The Interpreters' Newsletter*, *5*, 13–21.

Kurz, I., & Pöchhacker, F. (1995). Quality in TV interpreting. *Translatio-Nouvelles de La FIT-FIT Newsletter*, *15*(3), 4.

Mead, P. (2000). Control of pauses by trainee interpreters in their A and B languages. *The Interpreters' Newsletter*, *10*(200), 89–102.

Monti, C., Bendazzoli, C., Sandrelli, A., & Russo, M. (2005). Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus). *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *50*(4). https://doi.org/10.7202/019850ar

Moser, P. (1996). Expectations of users of conference interpretation. *Interpreting*, *1*(2), 145–178. https://doi.org/10.1075/intp.1.2.01mos

Ouyang, Q. (2020). Effects of non-verbal paralanguage capturing on meaning transfer in consecutive interpreting | Semantic Scholar. In *Multimodal Approaches to Chinese-English Translation and Interpreting* (1st ed., pp. 198–218). Routledge.

Pan, J. (2019). The Chinese/English Political Interpreting Corpus (CEPIC): A New Electronic Resource for Translators and Interpreters. *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, 82–88. https://doi.org/10.26615/issn.2683-0078.2019_010

Pietrandrea, P., Kahane, S., Lacheret-Dujour, A., & Sabio, F. (2014a). The notion of sentence and other discourse units in corpus annotation. *Spoken Corpora and Linguistic Studies*, 331–364.

Pietrandrea, P., Kahane, S., Lacheret-Dujour, A., & Sabio, F. (2014b). *The notion of sentence and other discourse units in corpus annotation*. Amsterdam/Philadelphia: John Benjamins.

Plevoets, K., & Defrancq, B. (2016). The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. *Translation and Interpreting Studies*,

11(2), 202–224. https://doi.org/10.1075/tis.11.2.04ple

Pöchhacker, F., & Zwischenberger, C. (2010). Survey on quality and role: Conference interpreters' expectations and self-perceptions. *AIIC Communicate! Spring*, *53*. http://aiic.net/p/3405

Poyatos, F. (2002). *Nonverbal Communication across Disciplines: Volume 1: Culture, sensory interaction, speech, conversation*. John Benjamins Publishing.

Rognes, T. (2001). ParAlign: A parallel sequence alignment algorithm for rapid and sensitive database searches. *Nucleic Acids Research*, *29*(7), 1647–1652.

Russo, M. (2018). Speaking patterns and gender in the European parliament interpreting corpus: A quantitative study as a premise for qualitative investigations. In *Making Way in Corpus-based Interpreting Studies* (pp. 115–131). Springer.

Russo, M., Bendazzoli, C., & Defrancq, B. (2018). *Making way in corpus-based interpreting studies*. Springer.

Russo, M., Bendazzoli, C., Sandrelli, A., & Spinolo, N. (2012). The European parliament interpreting corpus (EPIC): Implementation and developments. In *Breaking ground in corpus-based interpreting studies* (pp. 53–90). Springer.

Sandrelli, A. (2012). Introducing FOOTIE (Football in Europe): Simultaneous interpreting in football press conferences. In *Breaking ground in corpus-based Interpreting Studies* (pp. 119–153). Springer.

Sandrelli, A., & Bendazzoli, C. (2005). Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus). *Proceedings from the Corpus Linguistics Conference Series*, *1*.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora* (pp. 13–25). Springer.

Schmidt, T., & Wörner, K. (2009). EXMARaLDA–Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, *19*(4), 565–582.

Setton, R. (2011). Corpus-based interpreting studies (CIS): Overview and prospects. *Corpus-Based Translation Studies: Research and Applications*, 33–75.

Shlesinger, M. (1998). Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies. *Meta: Journal Des Traducteurs / Meta: Translators' Journal*, *43*(4), 486–493. https://doi.org/10.7202/004136ar

Stachowiak-Szymczak, K. (2019). *Eye movements and gestures in simultaneous and consecutive interpreting*. Springer.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In *Planning and task performance in a second language* (pp. 239–273). John Benjamins.

Thompson, B., & Koehn, P. (2020). *Exploiting Sentence Order in Document Alignment* (arXiv:2004.14523). arXiv. https://doi.org/10.48550/arXiv.2004.14523

Thormundsson, B. (2021). *Speech-to-Text transcript accuracy rate among leading companies worldwide in 2021*. https://www.statista.com/statistics/1133833/speech-to-text-transcript-accuracy-rate-among-leading-companies/

Tiedemann, J. (2011). *Bitext alignment* (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-031-02142-8

Tiselius, E., & Sneed, K. (2020). Gaze and eye movement in dialogue interpreting: An eye-tracking study. *Bilingualism: Language and Cognition*, *23*(4), 780–787.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2007). Parallel corpora for medium density languages. *Amsterdam Studies In the Theory and History Of Linguistic Science Series 4*, *292*, 247.

Vondřička, P. (2014). Aligning parallel texts with InterText. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1875–1879.

Wang, B. (2012). A Descriptive Study of Norms in Interpreting: Based on the Chinese-English Consecutive Interpreting Corpus of Chinese Premier Press Conferences. *Meta*, *57*(1), 198–212. https://doi.org/10.7202/1012749ar

Wang, B., & Li, T. (2015). An empirical study of pauses in Chinese-English simultaneous interpreting. *Perspectives*, *23*(1), 124–142. https://doi.org/10.1080/0907676X.2014.948885

Wang, B., & Zou, B. (2018). Exploring language specificity as a variable in Chinese-English interpreting. A corpus-based investigation. In *Making way in corpus-based interpreting studies* (pp. 65–82). Springer.

Wang, X., & Wang, B. (2022). Identifying fluency parameters for a machine-learning-based automated interpreting assessment system. *Perspectives*, 1–17. https://doi.org/10.1080/0907676X.2022.2133618

Yang, L. (2018). Effects of three tasks on interpreting fluency. *The Interpreter and Translator Trainer*, *12*(4), 423–443. https://doi.org/10.1080/1750399X.2018.1540211

Appendix A

Assessment criteria for information fidelity

Accuracy	Very Good (70-100)	Good (60-69)	Pass (50-59)	Poor (40-49)	Very Poor (0-39)
• Overall message • Secondary data • Omission • Distortion	Overall message present with little to no distortion or omission. Secondary and primary data retained accurately.	Meaning and message conveyed accurately with only some minor distortions and omissions.	Evidence of ability to analyse source text. Minor distortions evident but overall message conveyed accurately. Omissions present, but not undermining the global message.	Lack of ability to analyse leading to major distortions and/or *contresens* of global message or frequent distortions throughout the performance.	Poor analytical skills and/or comprehension leading to major distortions and/or *contresens* of the whole story. Ideas explained poorly resulting in a highly inaccurate rendition.

Appendix B

An Overview of the Corpora Constructed in this Study

Sub-corpus	Total token count	% of the entire corpora
The Multimodal Corpus of English-Chinese Conference Interpreter Training (MCECCIT)	87,066 source texts: 34,609 target texts: 52,457	44.7%
The Multimodal Corpus of Chinese-English Conference Interpreter Training (MCCECIT)	107,659 source texts: 65,746 target texts: 41,913	55.3%
Total	194,725	100%