Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

TRADUZIONE, INTERPRETAZIONE E INTERCULTURALITÀ

Ciclo XXXIII

Settore Concorsuale: 10/L1 Lingue, Letterature e Culture, Inglese

Settore Scientifico Disciplinare: L-LIN/12 Lingua e Traduzione, Lingua Inglese

AUTOMATIC SPEECH RECOGNITION (ASR) AND NMT FOR INTERLINGUAL AND INTRALINGUAL COMMUNICATION:
Speech to Text Technology for Live Subtitling and Accessibility.

Presentata da: Alessandro Gregori

Coordinatore Dottorato
Prof.ssa Raffaella Baccolini

Supervisore
Prof. Adriano Ferraresi

Co-supervisore
Prof.ssa Rachele Antonini

Esame finale anno 2021

Abstract (English)

Considered the increasing demand for institutional translation and the multilingualism of population in public space across Italy and Europe, the application of Artificial Intelligence (AI) technologies in multilingual communications and for the purposes of accessibility has become an important element in the production of translation and interpreting services (Zetzsche, 2019). In particular, the widespread usage of Automatic Speech Recognition (ASR) and Neural Machine Translation (NMT) technology represents a significant, recent development in the attempt of satisfying the increasing demand for interinstitutional, multilingual communications at inter-governmental level (Maslias, 2017). Recently, researchers have been calling for a universalistic view of media and conference accessibility which concerns not only persons with sensory disabilities (Greco, 2016), but anyone who have hearing difficulties in accessing audiovisual or speech content, as also indicated by the EU Audiovisual Media Services Directive (2016) and the latest international standards on subtitling, ISO/IEC DIS 20071-23 (Standardization, 2018). Given the frequent non-availability of interpreting human resources in international institutions for any language combination and at each single institution (Kalina, 2000), the application of ASR technology, combined with Neural Machine Translation, may allow for the breaking down of communication barriers between single speakers or among more individuals at European public conferences or public spaces, where multilingualism represents a fundamental pillar of institutional translation/interpreting (Jopek Bosiacka, 2013). In addition to representing a so-called disruptive technology (Accipio Consulting, 2006), ASR technology may facilitate the communication with people with minor hearing difficulties and with non-hearing users (Lewis, 2015). Thanks to Speech to Text technology, it is in fact possible to guarantee content accessibility for non-hearing audience via subtitles at institutionally-held conferences or speeches. Hence the need in this study for analysing and evaluating ASR output emerges: a quantitative approach is adopted to try to make an evaluation of subtitles output generated by ASR, with the objective of assessing its accuracy (Romero-Fresco, 2011). A database of F.A.O.'s and other international institutions' English-language speeches and conferences on the impact of Climate Change on the Agricultural Production is taken into consideration, which is analysed by applying a statistical approach based on WER and NER models (Romero-Fresco, 2016). The ASR software solution implemented into the study will be VoxSigma by Vocapia Research and Google Speech Recognition (via Descript/YouTube app). After having defined a taxonomic scheme, Native and Non-Native subtitles are compared to gold standard transcriptions. The intralingual and interlingual output generated by NMT is specifically analysed and evaluated in order to verify if ASR technology can be a valuable instrument to cope with the issues of communications with non-hearing persons at international institutions and spaces.

“For millions of years, mankind lived just like the animals. Then something happened which unleashed the power of our imagination. We learned to talk and we learned to listen. Speech has allowed the communication of ideas, enabling human beings to work together to build the impossible. Mankind's greatest achievements have come about by talking, and its greatest failures by not talking. It doesn't have to be like this. Our greatest hopes could become reality in the future. With the technology at our disposal, the possibilities are unbounded. All we need to do is make sure we keep talking.”

Stephen Hawking

Acknowledgements

I would like to dedicate this space of my PhD thesis to the persons who have contributed to the realization of it with their continuous support.

In primis, I would like to express sincere gratitude to my Supervisor, prof. A. Ferraresi, and to my Co-Supervisor, prof. R. Antonini, for their immense patience, their indispensable advise, the knowledge transmitted to me during the entire PhD thesis process.

I thank infinitely my parents and sister, who have always supported me and backed my any decision, since the very start of my academic career.

Heartfelt thanks to my colleague, Anna Zingaro, with whom I have shared the PhD programme. Thanks to her support and advise, I could cope with the difficulties encountered during the path.

I would also like to express my sincere thanks to Prof. P. Romero-Fresco (University of Vigo) and to Prof. H. Dawson (University of Roehampton), who have contributed to my project providing useful hints and material of research for my project.

Finally, I dedicate this PhD thesis to myself, my sacrifices and tenacity which allowed me to achieve this goal.

Contents

Abbreviation List

AI: Artificial Intelligence

API: Application Programming Interface

ASR: Automatic Speech Recognition

AST: Automatic Speech Translation

AT: Augmented Terminology

AVIDICUS: Assessment of Videoconference Interpreting in the Criminal Justice Service

AVT: Audiovisual Translation

BLEU: Bilingual Evaluation Understudy

CAI: Computer-Assisted Interpreting

CAT: Computer-Aided Translation

CL: Computational Linguistics

CORDIS: Community Research and Development Information Service

CRPD: Convention on the Rights of Persons with Disabilities

DARPA-GALE: Global Autonomous Language Exploitation of the U.S.' Defense Advanced Research Projects Agency

DNN: Deep Neural Network

DTW: Dynamic Time Warping

EBMT: Example-Based Machine Translation

ELF: English as Lingua Franca

EP: European Parliament

EPPS: European Parliament Plenary Session

EU: European Union

EU-BRIDGE: EU-Bridges Across the Language Divide

FAHQT: Fully Automatic High-Quality Translation

FAO: Food and Agriculture Organization of the United Nations

GSR: Google Speech Recognition engine

HAMT: Human-Aided Machine Translation

HHM: Hidden Markov Model

ICT: Information and Communication Technology

IT: Information Technology

LSTM: Long Short-Term Memory

LVCSR: Large Vocabulary Continuous Speech Recognition

MA: Media Accessibility

MAHT: Machine-Aided Human Translation

MI: Machine Interpreting

MT: Machine Translation

NER: Number of Edition and Error Rate

NLP: Natural Language Processing

NMT: Neural Machine Translation

NN: Neural Network

OOV: Out of Vocabulary

PBMT: Phrase-Based Machine Translation

RBMT: Rule-Based Machine Translation

RI: Remote Interpreting

RNN: Recurrent Neural Network

RQ: Research Question

SaaS: Software as a Service

SLT: Spoken Language Translation

SMT: Statistical Machine Translation

STS: Speech to Speech technology

STT: Speech to Text technology

TC-STAR: Technology and Corpora for Speech to Speech Translation

TER: Translation Error Rate

UDHR: Universal Declaration of Human Rights

UN: United Nations

VXS: VoxSigma suite, property of Vocapia Research

WER: Word Error Rate

1. Introduction

1.1. Preliminary considerations and hypotheses

The right to accessibility and to media accessibility are pivotal concepts for all accessibility studies and projects, as defined in Greco (2016: 1) and in Romero-Fresco (2018: 188). The concept of accessibility as a universalistic right stemmed from several regulatory initiatives by the United Nations' and the European Union's institutions in the course of the Twentieth and Twenty-First centuries. More specifically, the very concept of accessibility derives from the Universal Declaration of Human Rights (UDHR) of the United Nations (Paris, 1948) where the related concepts of *"human dignity"* and *"access"* were established for the first time. According to the UDHR, the concept of human dignity sets a minimum standard of quality of life (*i.e.*, essential resources for living) to which any individual is entitled for the sole reason of being a human being. On the other hand, the concept of *"access"* establishes the right to access to the essential resources required for a minimum standard of quality of life. In 1999, the UN Committee on Economic, Social and Cultural Rights contributed to better define this concept by also highlighting the role of each single State or national Government: *"the State must pro-actively engage in activities intended to strengthen people's access to and utilization of"* (p. 5 of the E/C.12/1999/5. General Comments) the objects of human rights (*i.e.*, the resources). More recently, the right of accessibility was certainly spurred by the approval of the UN *Convention on the Rights of Persons with Disabilities* (CRPD) of 2006. In particular, the *General Comment on Article 9* of the CRPD – released by the UN Committee on the Rights of Persons with Disabilities in 2014 – represents a milestone in the international disability movement to establish a new interpretation of disability and of persons with disabilities within society. The meaning of "disability" in fact shifted with changes in public policy. With the advent of universal civil rights protection against disability discrimination, what was addressed was not only whether the functionally compromised person is severely disabled enough to exercise a right, but whether mitigating interventions and reasonable resources can together achieve equitable access for that person. And this is of significant relevance for accessibility and media accessibility studies. Quoting Greco (2016: 2), *"assessing whether accessibility is a human right per se (or if not, then defining what exactly it is) is of the* *utmost importance for the field of human rights, as well as the struggle for inclusion of persons with disabilities*".

In recent years, the application of Artificial Intelligence (AI) Technologies has become an important element in the production of translation and interpreting services (Zetzsche, 2019), paving the way for further consolidation of (media) accessibility. In particular, the widespread usage of Automatic Speech Recognition (ASR) technology and Neural Machine Translation (NMT) represents a significant, recent development in the attempt of satisfying the increasing demand for interpreting and translation services at an interinstitutional and inter-governmental level (Maslias, 2017), not only in the EU, but also globally. Given the frequent, non-availability of interpreting human resources at the institutional level for any language combination and at each single institution (for example, see the work by Kalina, 2000 on legal and court interpreting), the application of Automatic Speech Recognition technology (namely, Speech to Text or Text to Speech technology), combined with Neural Machine Translation, may contribute to partially satisfy the demand and it may possibly help in breaking down the barriers of communication between the different EU countries or globally within the institutional context, where multilingualism certainly represents a fundamental pillar of Institutional Translation (Jopek Bosiacka, 2013). As a matter of fact, during the last decade, the scientific and academic debate on the usage of Automatic Speech Recognition (ASR) and Automatic Speech Translation (AST) technology (based on Neural Machine Translation) has significantly grown, together with the development of new ASR and NMT technologies, both at an academic level and at the level of international organizations. It is thus possible to maintain that the application of AI or AI-assisted technologies in the context of Institutional Translation/Interpretation has become an important element in the production of translation and interpreting services (as indicated by Alhawiti, 2015: 1439).

In connection to the instances of accessibility and to the right to media accessibility, while representing a so-called disruptive technology (Accipio Consulting, 2006), ASR technology should also be taken into consideration as it can facilitate the communication with non-hearing (deaf) users or final users with a partial hearing loss (Lewis, 2015), becoming an important tool for facilitating the communication in today's society, where the increasing ageing of the population is often synonymous with an increased number of hearing difficulties (see, for example

Goman, 2017). As a matter of fact, thanks to Speech to Text technology (and the production of live subtitles), it would be possible to guarantee content accessibility for non-hearing audience at institutionally-held conferences or speeches.

Starting from these preliminary considerations, the need for analysing and evaluating the Automatic Speech Recognition (ASR) and Neural Machine Translation (NMT) output emerges, together with a series of hypotheses which are identified in this introduction. The first hypothesis is that ASR technology can help in breaking down the barriers of communication at an institutional level at public conferences on specific scientific topic, within a multilingual context. Secondly, it is hypothesized that the combination of ASR technology with NMT may be fruitfully applied in the context of international organizations' debates, making it possible for them to cope with their accessibility needs. In particular, these technologies might produce live subtitles for breaking down the barriers of communication with non-hearing people or with individuals with minor hearing difficulties. Finally, the third hypothesis derives from the consideration that terminology plays an important role within the international organizations' debates, as documented in several works (see, for example, Cockhaert and Steurs. 2015). According to this consideration, it is here hypothesized that specialized (domain-related) terminology should be further investigated for accessing its impact on subtitles quality.

To sum up the preliminary considerations above, it is important to point out that this study originates from a combination of different needs. First of all, the need for meeting the increasing demand of translation and interpreting services at conferences at an international, institutional level (also the necessity of reducing expense costs in institutions' budgets). Secondly, the necessity of responding to the accessibility requirements provided by the recent EU Directive on Audiovisual and Media Services (2016); thirdly, the widespread usage of ASR technology in combination with Neural Machine Translation (NMT) poses a series of challenges which, to my knowledge, were not probably sufficiently examined in the scientific literature within the specific context and the communication scenario presented here; fourthly, the importance of terminology in Institutional Translation generates further considerations in connection with the usage of ASR and NMT technologies at international organizations such as the European Union, the United Nations and the Food and Agriculture Organization (FAO). Hence it is possible to assert that the literature framework for this study should be grounded on four main pillars of discussion and study: 1. Automatic Speech Recognition theories and studies; 2. Neural Machine Translation theory and studies; 3. Accessibility Studies; 4. Institutional Translation theory and studies. These four branches of scientific knowledge are fundamental to this study and they will be better described in Chapter 2 dedicated to the literature review.

As far as the organization and structure of contents are concerned, this study will firstly present the main studies and theory on Automatic Speech Recognition and Neural Machine Translation, with the intention of conducting a critical review of the works that are more relevant to the present study, with a special focus on those where an evaluation of accuracy was conducted. This will be accompanied by an in-depth analysis of the technological evolution of these technologies with the objective of grasping the main requisites for an effective ASR and NMT-based system capable of providing quality output for live subtitling (Chapter 2). After this literature and technology review, a definition of the methodological approach adopted in this study will be described in detail, with the objective of defining a series of Research Questions (Chapter 3) and setting up a methodology for the subsequent processing of data and the configuration of an experimental pipeline for the implementation of the ASR and NMT technologies within the institutional context. In Chapter 3, a taxonomic scheme will also be offered in order to establish a categorization of ASR and NMT errors for the subsequent evaluation of accuracy, including the definition of an effective instrument for the validation of the taxonomic scheme (*i.e.*, inter-annotator agreement). Chapter 3 will finally offer and describe the statistical models used for the analysis of accuracy, while highlighting the weaknesses and strengths of the statistical models proposed by other scholars (*i.e.*, the WER and NER models). After having defined an appropriate methodology for the experimental part of the study, it will then be possible to carry out an analysis of data in Chapter 4, focusing on the validation of the taxonomic scheme defined here and on the evaluation of accuracy, both for the Automatic Speech Recognition and for the Neural Machine Translation outputs. The evaluation will be conducted according to different instances of analysis, by taking into account the diversified needs of final users and the different application scope: namely, intralingual subtitling (in English) and interlingual subtitling (in Italian) for non-hearing people or for users with a partial loss of hearing.

At this stage of the introduction, before discussing about the topics described in the paragraph above, it is important to describe the specific context and the communication scenario developing around the object of the next analysis.

1.2. The communication scenario

The communication scenario of the present thesis is represented by public conferences on climate change held at international organizations or public institutions. The decision of choosing an institutional setting for the present study is based on the idea that institutional organizations can offer a multilingual context where the principles of diversity, linguistic identity and accessibility can effectively coexist. In particular, the selection of institutions such as the European Union and the Food and Agriculture Organization (FAO) was made by considering these institutions' policies in favour of accessibility and multilingualism. Additionally, these institutions can offer a plethora of audio/video materials that are easily consultable and open to the public domain (their use does not require any authorization). The actors of the present scenario are the speaker (a representative, political leader or an expert/scientific scholar in the field of climate change) and the target audience (consisting in the international organization's members, experts in the field of climate change, stakeholders, citizens, etc, with a particular focus on dear or hard of hearing people). The communication scenario is described in detail in the database of the present thesis (Chapter 3), where international organizations' English-language speeches on the impact of climate change and its effects on agriculture (available in Appendix A) are collected, together with an analysis of a dataset based on statistical, quantitative models. More specifically, the software solutions implemented in the present study to transcribe those speeches will be *VoxSigma* suite, property of Vocapia Research, and *Google Speech Translation* engine technology, to be deployed via *YouTube* or *Descript* applications (both apps are used as SaaS solutions: *i.e.* "Software as a Service"). In Chapter 3, the criteria for the selection of these technologies will be better clarified. In general, it is possible to highlight that a speech-to-text pipeline includes the following set of technologies, which consists of three primary components: A. the Automated Speech Recognition (ASR) technology; B. the Machine Translation (MT) engine; and, finally, C. the Speech to Text (STT) component. As described by Lewis:

> *"The first, ASR, converts an input audio signal into text, essentially "transcribing" the spoken words into written words. (...) Machine Translation (MT), the second component, maps words and phrases in one language to words and phrases in the second" (Lewis, 2015: 59).*

Under this pipeline, which is going to be described in more detail in Chapter 3, the STT (Speech to Text) component finally converts into text (or subtitles) the original source input. To better clarify what SST is, it should be added that under the present study, the STT output is coincident with the subtitles generated by NMT in the target language: Italian. In other studies from the reviewed literature, it is possible to find an STS (Speech to Speech) component as well, which requires a Speech or Voice Synthesizer component to reproduce the NMT output by voice. Given that the aim of the present study is that of examining the accessibility of content in the form of subtitles, the STS component will not be considered. Finally, it should be anticipated here that the STT output will follow the segmentation provided by default by the ASR solution implemented.

From a general perspective, in today's societies, one may be apt to think of speech, and language, as a writing system that may be pronounced. In point of fact, as reported by Crystal and Robins (2020), *"language generally begins as a system of spoken communication that may be represented in various ways in writing"*. Without entering into a description of the physiological aspects and anatomic nature of speech production, it is here sufficient to mention the definition offered by Crystal and Robins to describe "speech" and the act of speaking:

> *"Speaking is in essence the by-product of a necessary bodily process, the expulsion from the lungs of air charged with carbon dioxide after it has fulfilled its function in respiration. Most of the time one breathes out silently, but it is possible, by adopting various postures and by making various movements within the vocal tract, to interfere with the egressive airstream so as to generate noises of different sorts. This is what speech is made of." (Crystal and Robins, 2020)*

At this point, the communicative context in which the ASR process is reproduced should be presented to better understand and identify the role and function of ASR in a speech production and recognition process. From a sociological and psychological perspective, normally, when a speech takes place between two individuals or between a speaker and its audience, a form of communication is carried out. According to Gordon (2020), *"communication is the exchange of meanings between individuals through a common system of symbols"*. In linguistics, as explained by Gordon (2020), this event of communication is developed according to a psycho-linguistic linear model containing five elements: *i.e.*, an information source, a transmitter, a channel of transmission, a receiver, and a destination, all arranged in linear order. With the introduction of the electronic format for messages and communication and the expansion of multilingualism, this linear model was modified to include six components: (1) a source, (2) an encoder, (3) a message, (4) a channel, (5) a decoder, and (6) a receiver (Gordon, 2020). In a simpler way, when describing the speech process in full, Lewis (2015) describes the speech development process by indicating the following ones as the fundamental steps in a communication situation: 1. the speaker formulates his/her ideas into words; 2. the speaker generates sound using the vocal cords and speech system; 3. sound is transmitted via an acoustic wave in air to the ear of the listener as vibrations; 4. sound is transmitted to the listener's brain via auditory nerves; 5. those vibrations are converted to some "language" in his/her brain; 6. the brain extracts meaning from sound. This simple, basic process for communication, which can be applied to all conversation and speech situations, is also represented in Figure 1.1 below, where the different steps described above are included.

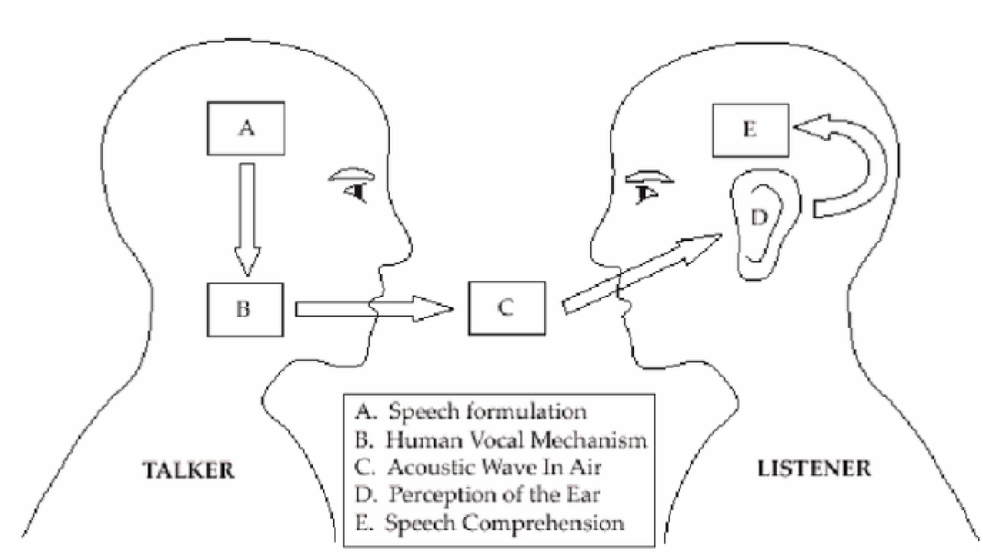


Figure 1.1 - Diagram of speech production and perception process (Towards Data Science, 2019).

After these preliminary considerations, it is therefore possible to specify that the object of processing for the ASR technology should be identified with the element C in Figure 1.1 above, i.e., the "acoustic way in air" bearing the speech formulation signal or the message (to use the previous linear model). The role and function of the ASR system should therefore be located into this position of any communication process. Finally, it is necessary to point out the clear limitations of this basic model of communication, as the fundamental role of non-verbal and contextual cues are missing.

With reference to the institution setting of the communication scenario examined in the present study, it is important to highlight the function and role of subtitling and respeaking processes. These two services in fact contribute to the accessibility of content at public conferences and they are often accompanied with human interpreting or automatic translation services when the communication is from a source language to one or more target languages. In the present scenario, the source language is English and the target language is Italian. More specifically, it is necessary to comment that real-time or live subtitling within the institutional context can be displayed at public spaces in different modalities: on a screen behind the speaker, on a screen next to or far from the speaker, or directly on a remote screen if the audience is assisting to the conference in the remote modality (broadcasting service). Under any circumstances, all subtitles are reproduced in real time but with a certain latency with respect to the source speech. Another important consideration to be made about the breaking down of communication barriers at conferences is represented by the respeaking service which often supports the subtitling process in producing accessible content. The role of the respeaker is that of editing the subtitles when they are not sufficiently understandable to the target audience. The editing process is carried out in real time and it adds up to the subtitling process. It is therefore evident that respeaking contributes to increase latency in the end, though improving the accuracy of the service.

When discussing about the target audience of the communication scenario analysed here, it is first of all appropriate to make a distinction between "deaf" (or totally non-hearing people), "hard of hearing", "hearing impaired" and "deafened". In fact, this distinction will be particularly relevant when examining the final output of the communication scenario examined. In common use, there is often confusion over these terms, both in their definition and appropriateness of use. Generally speaking, the term "hearing impaired" is used when it is intended to describe people with any degree of hearing loss (from mild to profound), including those who are deaf and those who are hard of hearing. The term "hearing impaired" implies a deficit or handicap so people prefer to use the other words, which are considered more politically correct. When someone is deaf, he/she has hearing loss so severe that there is very little or no functional hearing. When people have a loss of their hearing ability, they are called as "hard of hearing"; with these persons there may be enough residual hearing that an auditory device, such as a hearing aid or FM system, provides adequate assistance to process speech. Finally, the "deafened" people are generally those individuals who become deaf as an adult and, therefore, may face different challenges than those of a person who became deaf at birth or as a child. In addition to using hearing aids, cochlear implants, and/or other assistive listening devices to boost available hearing, all the target audience of the present communication scenario may read lips, use sign language, sign language interpreters, and/or subtitling. In the present study, the two main categories of the target audience will be the deaf (also indicated as non-hearing or totally non-hearing people) and the hard of hearing (also referred to as people with hearing difficulties or with a partial loss of hearing).

During the last two decades, several international and national initiatives were conducted, both at a public and private levels, with the aim of investigating on the use of Artificial Intelligence (AI) for the breaking down of communication barrier and also for implementing the interpreting and translation services. In the present study, it should be specified that the scientific debate on the use of Automatic Speech Recognition (ASR) and on Neural Machine Translation (NMT) has significantly grown during the last decade, together with the development of new ASR and NMT technologies, both at an academic level and at the level of international organizations and institutions (namely, the European Union and the U.S. Government). The application of Artificial Intelligence (AI) or AI-assisted technologies in the context of Institutional Translation/Interpretation has *de facto* become an important element in the production of translation and interpreting services (Alhawiti, 2015: 1439). In particular, the widespread use of Automatic Speech Recognition (ASR) technology represents a significant, recent development in the attempt of satisfying the increasing demand for interpreting at an interinstitutional and inter-governmental level (as also commented by Maslias, 2017), not only in the EU but also globally. Given the frequent non-availability of qualified interpreting human resources at the institutional level for any language combination and at each single institution (see for example the work by Kalina, 2000), the application of ASR technology (namely, Speech to Text or Text to Speech technology), combined with Neural Machine Translation, may contribute to breaking down communication barriers between EU countries or globally, where multilingualism represents a fundamental pillar of institutional translation/interpreting (Jopek Bosiacka, 2013: 110). But, in addition to representing a so-called disruptive technology (Accipio Consulting, 2006: 30) for its impact on the interpreting/translation industry and on every people's life, it should also be noted and highlighted that ASR technology can facilitate communication with non-hearing users (Lewis, 2015: 58), or with users having hearing difficulties. As a matter of fact, thanks to Speech to Text technology (and the production of real-time or asynchronous intra- or interlingual subtitles), it is possible to allow accessibility for non-hearing audience at institutionally-held conferences or speeches. More specifically, the implementation of AI included the use of Automatic Speech Recognition (ASR) and/or Neural Machine Translation (NMT) for public services or institutions, namely in the United States and within a plethora of different European Institutions and academic institutions. In this respect, it should be observed that, common to most of previous projects is a pipeline (partially similar to the present study's pipeline – see Chapter 3 for further details), which develops into three main steps as in Figure 1.2 below.

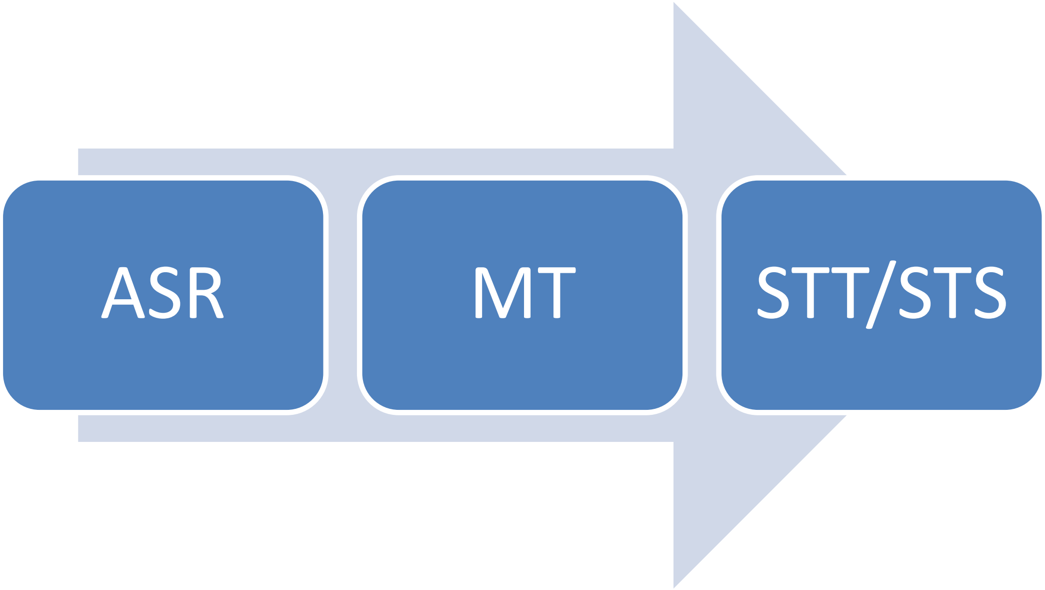


Figure 1.2 - Common basic pipeline implemented in previous research projects on ASR.

Across this pipeline, to re-quote Lewis: *"ASR (Automatic Speech Recognition) first converts an input audio signal into text, essentially "transcribing" the spoken words into written words"* (Lewis, 2015: 59). Then Machine Translation (MT), the second component in Figure 1.2 above, maps the words and phrases in one language to words and phrases in the second target language. As we will see more in detail in the section dedicated to Machine Translation technology in Chapter 2, MT may incorporate a statistical-based model (SMT) or a neural network model (NMT), or even a combination of both models. Finally, at the end of the process, the Speech to Text (STT) component maps text in a given language to a text form, and is generally trained on carefully recorded audio and transcripts from one native speaker. This pipeline was also at the basis of several research projects conducted in recent years, though with differences in the components combination and analysis methodology. Many of these projects are described in Chapter 2, §2.2.3.2 (Verbmobil, TC-Star, DARPA-Gale, EU-Bridge), and they incorporate several subprojects and research outcomes. At this stage, it is enough to underline that, like in the present study, the previous projects examined here were based on an automatic pipeline where human intervention was limited to a minimum. In fact, ASR technology is accompanied with MT but also in combination with interpreters or the intervention of respeakers for the production of live subtitles (semi-automatic workflows). Yet it should be remarked that the evaluation system adopted in the previous projects present a series of limitations which are going to be discussed in Chapter 2. In general, it is possible to assert that, with respect to the methodology, the limitations connected with qualitative instruments (questionnaires), and the risk of subjectivity, as well as those connected with the statistical measures implemented did not allow to measure accuracy of these technologies within an institutional context like the one examined here. Additionally, the present study is based on four main branches of knowledge, as it is possible to see in Chapter 2, while previous works did not attempt to combine the different disciplines around ASR. Finally, as we will see in the next chapters, the present study should also be considered innovative in presenting and examining the impact of terminology in the output quality evaluation and also in evaluating a specific topic across international organizations' debates: i.e., "climate change".

1.3. Summing up

In this introduction, a series of general considerations were expressed in order to understand the need for a study on the combination of ASR and NMT technologies. After having presented the instances and needs of accessibility and institutional translation in today's society, and in particular, across the international organizations where multilingualism represents a fundamental pillar of their identity, a few hypotheses were defined for the purposes of this study. As mentioned in Section 1.1. above, the effectiveness of previous projects and international initiatives in defining a methodology and a processing pipeline combining ASR and NMT could be improved. In particular, to assess the initial hypotheses of this study, it is necessary to identify and verify valuable metrics and instruments for the evaluation of the final output and its accessibility. Additionally, the current evolution of today's technology (both ASR and NMT) urges an in-depth review of the state of the art, both from a technological point of view and from a scientific point of view. For this reason, in Chapter 2, a review of literature and the state of the art of ASR and NMT technologies will be conducted to better identify the criticalities and potential possibilities of improvement in the definition of an effective methodology and selection criteria for the identification of the most advanced ASR and NMT solutions, including the identification of the most suitable tools and protocols for an effective evaluation of accuracy.

2. Literature Review

2.1. Introduction

This chapter on the Literature Review intends to present and describe the multifaceted theoretical background of the present study. In the attempt of creating a general framework for the different disciplines which are relevant to this study, a wide array of works are taken into consideration. These works belong to specific disciplines which will be represented here as "drawers" from which it is possible to draw useful materials for an appropriate consolidation of the theoretical framework. More specifically, the disciplines treated under this review are: 1. the theory and studies on Automatic Speech Recognition; 2. the theory and studies on Neural Machine Translation; 3. Accessibility Studies; and, 4. Institutional Translation. These disciplines exist *per se* but, to my knowledge, no other study has tried to combine them in a study. The "weight" of each discipline varies according to the relevance of the object to this thesis. For this reason, with the intention of representing the interrelation and interoperability of each discipline with respect to the others, and their relevant significance, a figure is created to describe all that in a graphic form (see Figure 2.1 in the next page). To comment on Figure 2.1 below, it is possible to see that ASR and Accessibility Studies are the most important disciplines for the present study (given the role of ASR technology and the objective of accessibility) and all the four areas of studies are strictly interconnected with each other, though it should be pointed out that this study will mostly be based on ASR, NMT and Accessibilities Studies and, to a minor extent, on Institutional Translation. Together, all these disciplines converge to create a framework for the entire study in an innovative way. In this respect, it should be underlined that the four areas explored as a starting point for the literature review do not pertain to the same "level" in the scientific literature, though they are here treated as being on the same level. In fact, Accessibility Studies is a discipline in itself, while Institutional Translation identifies a specific type of translation so possibly it represents a sub-field of Translation Studies, rather than a disciplinary area *per se*. Similarly, ASR (as explained in §2.2.1) may both refer to a disciplinary area of research and to a technological system and process (as it also happens with NMT).

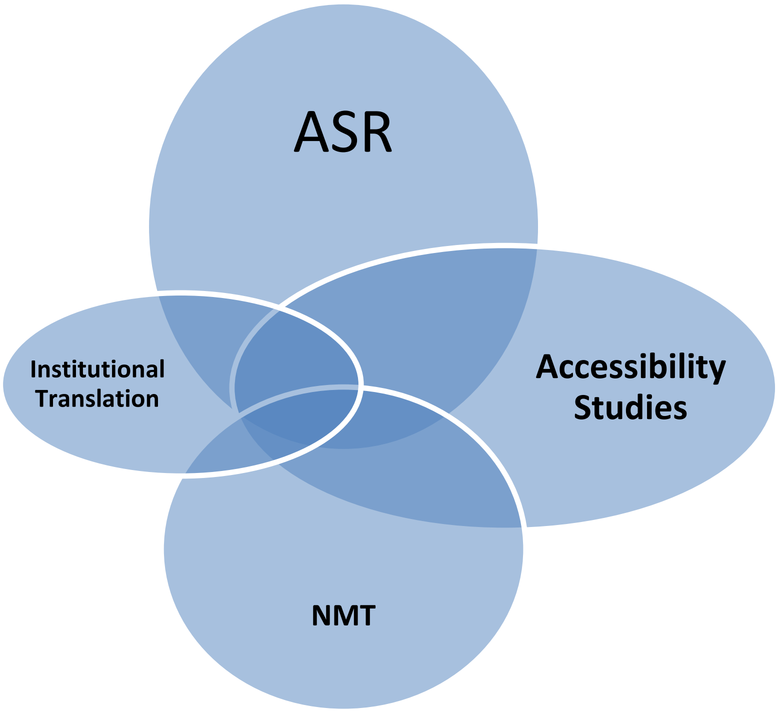


Figure 2.1 - Representation of the interoperability of the 4 scientific disciplines.

For an overview of its content organization and structure, this chapter will present the literature framework about the studies and theory on Automatic Speech Recognition (ASR) (§2.2) and Machine Translation (MT) (§2.3) in order to provide for substantial background knowledge and also scientific grounds for the definition of a methodology (which is better described in Chapter 3), as well as for the development of this study's analysis (Chapter 4). In particular, for a general but not exhaustive literature review, a history of Automatic Speech Recognition (ASR) and Machine Translation (NMT) technology is offered to better understand the technological development and evolution leading to current state-of-the-art technology (see §2.2.2 and §2.3.2 below). Provided that this study aims (see §3.2 on Research Questions on Chapter 3) to evaluate the potential and the role of ASR technology in breaking down the barriers of communication for non-hearing people and for accessibility purposes, a general presentation of research studies on Accessibility and Media Accessibility is also offered in §2.4. Finally, considered that this study focuses on the performances and the role of ASR technology for the generation of subtitles for live conferences or speeches held at institutional organizations, the function of Institutional Translation and its current development is discussed together with a presentation of recent studies on the role of Artificial Intelligence (AI) in institutional interpreting and translation services (§2.5).

2.2. Studies and theory on Automatic Speech Recognition

Under this section, a definition of Automatic Speech Recognition (ASR) will be offered, together with a distinction between ASR and Automatic Speech Translation (AST) and/or other subfields of research. Additionally, the evolution of ASR technology will be presented in order to better understand the development of this technology and the current cutting-edge technology available in the market, with a special focus on the architecture typologies at the basis of it. Finally, a critical analysis of the literature produced around ASR will be supplied with the objective of identifying the existing criticalities of previous studies and the potential areas of improvements for the purposes of this study.

2.2.1. Definition of Automatic Speech Recognition

When defining the concept of Automatic Speech Recognition, it is necessary to carry out a fundamental distinction between Automatic Speech Recognition (ASR) and Automatic Speech Translation (AST), as in literature there is often a combined usage of both (for example in Fügen et al., 2006; Lazzari, 2006; Matusov et al., 2008) or a not-so-clear delimitation of their scope (Romero-Fresco, 20018). To quote Fantinuoli and Prandi (2018: 169), AST is *"the technology that allows the translation of spoken words from one language to another by means of computer programs"*. More specifically, AST incorporates three technological components to perform the task: ASR, MT and STT. The first consideration to be done is therefore about the fact that ASR can be seen as a component of AST. When defining ASR, it should be underlined the fact that two different definitions are required when referring to it as a disciplinary area of research and as a technological system and process. More specifically, quoting Rabiner and Juang, ASR should be considered as *"an interdisciplinary subfield of computational linguistics capable of integrating several skills and an array of knowledge from more areas of studying"* (Rabiner and Juang, 1993) so as to develop methodologies and technologies allowing for the recognition and translation of the speech in a text by means of IT devices or computers. To my knowledge, apart from being denominated “automatic speech recognition” (ASR), in the scientific literature, this subfield of research is also indicated as “computer speech recognition”, or simply as “speech to text” (STT). In line with Rabiner and Juang, Maffi defines Automatic Speech Recognition as:

> “*…an interdisciplinary subfield of computational linguistics, at the crossroads between linguistics, computer science and electronics engineering, whose main aim is to develop methodologies and technologies allowing transcription of spoken language into text by using computer devices*”. *(Maffi, 2016: 17: my translation).*

Maffi’s definition certainly tries to describe the technological component of this interdisciplinary subfield of computational linguistics. In fact, Automatic Speech Recognition should also be defined as a technological system and process. From this perspective, to define ASR as a system or process, it is possible to use the words by Stuckless, who describes ASR as an *“automatic transcription of speech in real time in a readable text, a process by which human oral speech is recognized”* (Stuckeless, 1994: 197). On the other hand, Dureja and Gautam define ASR in combination with Machine Translation (as if ASR is always interconnected with MT, but this is not always the case): *“a process that takes the conversational speech phrase in one language as an input and translated speech phrases in another language as the output”* (Dureja and Guatam, 2015:28). During the last decade, a significant improvement was obtained in terms of ASR technology progress and performance. In fact, together with Kumar et al. (2015: 229), it is possible to highlight that *“over the past decade, considerable progress has been made in developing usable, two-way speech-to-speech (S2S) translation systems that enable real time cross-lingual spoken communication”*. And this idea also finds support in the words by Lewis:

> *“Although flawless communication using speech and translation technology is beyond the current state of the art, major improvements in these technologies over the past decade have brought us many steps closer”. (Lewis, 2015: 58).*

A further distinction should be made between Automatic Speech Recognition (ASR) and Speaker or Voice Recognition. As already mentioned above, ASR can be simply defined as the automatic recognition of a speech, while Speaker or Voice Recognition implies *"the recognition of the physical properties of a voice, to identify the speaker"* (Eugeni, 2008: 15; my translation) on the basis of a comparison between speech input data previously collected. This type of process (and technology) is generally used in domotics or in security systems for the identification of an individual, but it may also be incorporated into an advanced ASR system for the speaker recognition.

At this point of this review, after having defined Automatic Speech Recognition both as a discipline and as system or process, it is necessary to describe the history and evolution of this technology in order to better understand the scope, the architecture and the application of this technology.

2.2.2. History and development of ASR technology

Under this section, the main steps in the rise and development of ASR technology are presented in chronological order. Following a series of scientific preliminary investigations on speech recognition started in 1932, in 1952 the Bell Labs developed the first software for the speech recognition capable of recognizing numeric values spoken out by a speaker. Yet the period's technologies could not offer a voice recognition service for words recognition (Pierce, 1969). During the 1960s, a student from Stanford University, Raj Reddy, implemented the first system of Continuous Speech Recognition that required no pauses between a word and another. It was in those years that the design of Speech Recognition software moved from the usage of a dynamic time warping (DTW) system to the usage of Hidden Markov Models (HMM). For further technical details on DTW and HMM, it is possible to consult the definitions offered by Müller (2007: 69) and Eddy (2004: 1315), respectively.

Later on, thanks to computer hardware and software developed in the 1980s, IBM, under the direction of Fred Jelinek, invented a typing machine which was speech-driven by means of the first commercially-marketed dictation solution denominated *Tangora*. This solution was able to recognize a vocabulary of as many as 20,000 words. Yet this innovation was not able to offer Speech Recognition in rapid times. Given the limited RAM capacity of the typing machine, *Tangora* would take a lot of time to elaborate a few minutes of dictation (McKean, 1980).

At the beginnings of the following decade, thanks also to further developments in computer technology, new features were incorporated into the ASR technology, including speaker independence and resistance to noise features. Speaker independence permits the ASR software solution to recognize any speaker voice without the need for preventively training it to the speaker's voice, while resistance to noise consists of the possibility of isolating speech from background noise such as road traffic, or other disturbances. It was in the 1990s that Xuedong Huang, a student of Raj Reddy, created *Sphinx II*, the first Speech Recognition software capable of offering the new functionalities mentioned above (speaker independence and noise resistance).

However, the most important breakthrough in the design and development of ASR solutions came with the start of the new millennium. In fact, in the early 2000s, the statistical recognition model (*i.e.*, the Hidden Markov Model) was replaced by Neural Networks (NN) or Deep Neural Networks in the projecting of the ASR engine. Accuracy and speed were then improved thanks to the incorporation of these new systems. More specifically, Neural Networks represent *"an attractive acoustic modelling approach"* (Zahorian et al., 2002) in ASR. When used to estimate the probabilities of a speech feature segment, *"neural networks allow discriminative training in a natural and efficient manner"* (Karpagavalli and Chandra, 2016: 400). On the other hand, a Deep Neural Network (DNN) can be described as *"an artificial neural network with multiple hidden layers of units between the input and output layers"* (Hinton et al., 2012). Given the complex nature of these systems, it should be here enough to maintain that these models made a deeper recognition of signal possible thanks to the possibility of training the ASR solution. In simple words, "to train" the ASR system means to expand the ASR recognition capability by entering more and more reference materials (vocabulary or corpora of texts) into the system: the more the ASR solution is trained and expanded (with larger vocabularies), the more accurate it gets.

In recent years (and up to present day), ASR technology has rapidly developed reaching outstanding performances thanks to the combination of the statistical model (HMM) with neural networks (NNs). This has led to the design of a new particular neural network system denominated as Long Short-Term Memory (LSTM). As defined in Hochreiter and Schmidhuber (1997), the LSTM system is an artificial recurrent, neural network-based architecture. Unlike previous NN systems, LSTM has feedback connections, which means that it can "remember" connections already established before, but for a longer time, if compared to previous systems. Long Short Term Memory networks are in fact *"capable of learning long-term, recurrent dependencies"* (Colah's Blog, 2015). These systems have for example been incorporated into popular-across the market ASR solutions such as *Google Voice* (and Speech Recognition engine) in 2015 and, previously, in *Skype* (now Microsoft Skype Translator) in 2011. At the same time, in recent years, the arrival of high-speed Internet connection (2010s) and the possibility of creating cloud-based ASR technologies have added further improvements to the usability of this technology via Web interfaces (APIs) allowing users to leverage it from remote positions.

During the last decade, as already mentioned above, the most important innovation has been represented by the progressive development of Deep Neural Networks (or DNNs) and their incorporation into Automatic Speech Recognition technology. To put it in simpler words, the denomination of *deep neural networks* comes from the analogy with human brain's structures and with its functioning mechanism. In fact, like human neurons, which can operate on a separate, singular basis and which are connected with each other by means of axons and synapses, in the same way the (artificial) neural networks are composed of a high number of standalone, elaboration units (also called "neurons") which are interconnected. DNNs also include an algorithm which modifies the significance or "weight" of each single connection so that the input signal can be directed towards a determined direction and the processing can be oriented towards a given output. This technology is generally used in the execution of the so-called "pattern recognition", that is to say an elaboration of data made to create matchings between complex inputs and simple outputs. To better explain what is pattern recognition in a machine, it is possible to use the description offered by Fu to indicate the pattern recognition process elaborated by a human brain:

> *"The problem of pattern recognition usually denotes a discrimination or classification of a set of processes or events. The set of processes or events to be classified could be a set of physical objects or a set of mental states". (1976: 1-2)*

As commented in the definition above, human beings perform the task of pattern recognition in almost every instant of their image or data processing operations during their working lives. But in the last decade, pattern recognition also started to be performed by machines or computers thanks to the use of Artificial Intelligence. To easily understand the method used in pattern recognition by humans or computers, generally speaking, it is possible to quote Fu (ibid) again:

> *"the many different mathematical techniques used to solve pattern recognition problems may be grouped into two general approaches; namely, the decision-theoretic (or statistical) approach and the syntactic (or linguistic) approach."*

For the purposes of describing the pattern recognition process carried out by state-of-the-art ASR technology, it is necessary to specify that current technology makes use of both approaches defined by Fu above. In particular, in ASR, the objective in pattern recognition is *"to classify an unknown pattern as one from a set of candidate groups"*, as also commented in O'Shaughnessy (2008: 2968). In speech recognition, this implies labelling each input utterance (i.e., the audio input or sound waveform) with its corresponding text.

The final step in the evolution of ASR technology and architecture is based on a combined usage of HMMs (the statistical method) and Deep Neural Networks: this approach is different from previous technology as it does not eliminate the usage of HMMs, but it combines them with the usage of DNNs, offering a significant improvement in terms of ASR software performances (Sturari, 2012). The functioning mechanism is based on the two approaches defined above by Fu (namely, statistical and linguistic). In other words, the same neural network operation might generate different outputs with the same inputs. The way in which the input data are processed depends on the algorithm governing the network. More precisely, in order to produce the expected or intended output and results, it is necessary to "train" the network (as described above) and improve the algorithm by implementing the procedure described by Falletto:

> *"[...] a pulse is sent to the input of the network and the output generated is then observed. Afterwards, the weights of the connections so produced are modified so as to obtain an output which is closer to the output required or expected. Further inputs are sent in a series, the outputs so generated are then measured and the process is repeated as many times as it is necessary. A neural network, after the training phase, is capable of supplying a coherent output even if it receives an input which was not entered during the training phase". (Falletto, 2007: 64; my translation).*

Given their high flexibility, or capacity of adapting to new data and to the combination with the HMMs, DNNs are now largely implemented in ASR technology development. In fact, as commented by Beaufays, *"around 2012, Deep Neural Networks (DNNs) revolutionized the field of speech recognition"* (Beaufays, 2015). Thanks to the implementation of the most advanced hardware and software technologies, and to the access to big data, ASR systems can now access to data in a faster, easier way and to "learn" more rapidly. Currently, many ASR industry players (such as Google, Microsoft, IBM, Baidu, Apple, Amazon, Nuance, etc.) have developed a wide range of solutions based on this combination for their ASR system products. The introduction of the Cloud technology also improved the recent ASR technology implementation and the combined use of DNN and HMM architecture technology.

2.2.3. Studies on ASR

Extant literature on the studies and theory of Automatic Speech Recognition, as per the definition provided in §2.2.1. above can be grouped, to my knowledge, into four main groups of studies. In the first group, it is possible to find a series of studies mainly focusing on the technological aspects of this system and on its architecture definition; in the second group, it is possible to collocate the works describing the potential application and combination of ASR with other technologies, for example with Machine Translation or Voice Synthesis; the third group includes studies and works focusing on the combination of ASR with the interpreting service or with the work of interpreters in the booth; finally, in the fourth group, studies on Accessibility and on the production of subtitles for non-hearing people can be collected. More in particular, it is important to underline that the first group of studies is mainly based on the description of engineering or IT notions and knowledge, thus it requires advanced engineering or software programming skills for the comprehension of these notions. This group of studies will therefore be reviewed only for the purposes of defining the different architecture systems available for ASR, without entering into more details regarding engineering or software components. As far as the second group of studies is concerned, a general review of previous umbrella projects will be offered in order to point out the evolution of ASR combined with other technologies (namely, MT and NMT), including the methodologies and the metrics used for the purposes of evaluating the combined application of MT and ASR in the final step of quality analysis; thus, the focus of this part of the review will be to underline the main instruments and parameters for quality evaluation available in literature. The third group of studies will be reviewed for the purposes of underlining the potential advantages and criticalities relating to the combination of ASR with the interpreting service/work and, in particular, with the function of automatically translating oral material. Finally, the fourth group of studies will be examined and reviewed to better understand the background and potential contributions of previous works on ASR and Accessibility, with a particular reference to the production of subtitles for non-hearing people. This group of studies will be presented and described in the section dedicated to Accessibility Studies (§2.4) under this chapter.

2.2.3.1. Studies on the architecture and components of ASR technology

Starting from the first group of studies listed above, it should be commented that a plethora of works can be collected under it, including Anusuya and Katti (2011), De Watcher et al. (2007), Deng and Li (2013), Garofalo et al. (1993), Ghai and Singh (2012), Hemdal and Hughes (1967), Huang and Deng (2010), Huang et al. (2001), Jurafsky and Martin (2009), Karpagavalli and Chandra (2016), Li et al. (2014), Yu and Deng (2015), to list just some of the consulted works. Without entering into more details about the engineering and programming notions and knowledge offered into those works, the focus of this review is on the architecture and components characterizing ASR. In particular, it is certainly worth commencing with the presentation of the work by Karpagavalli and Chandra (2016), who carefully described the architecture and components of ASR, including the different architectural systems at the basis of it. According to these scholars, a ASR architecture can be defined as follows: *"A typical speech recognition system is developed with major components that include acoustic front-end, acoustic model, lexicon, language model and decoder"* (Karpagavalli and Chandra, 2016: p.394). Figure 2.2 below clearly shows the components of a typical ASR architecture.

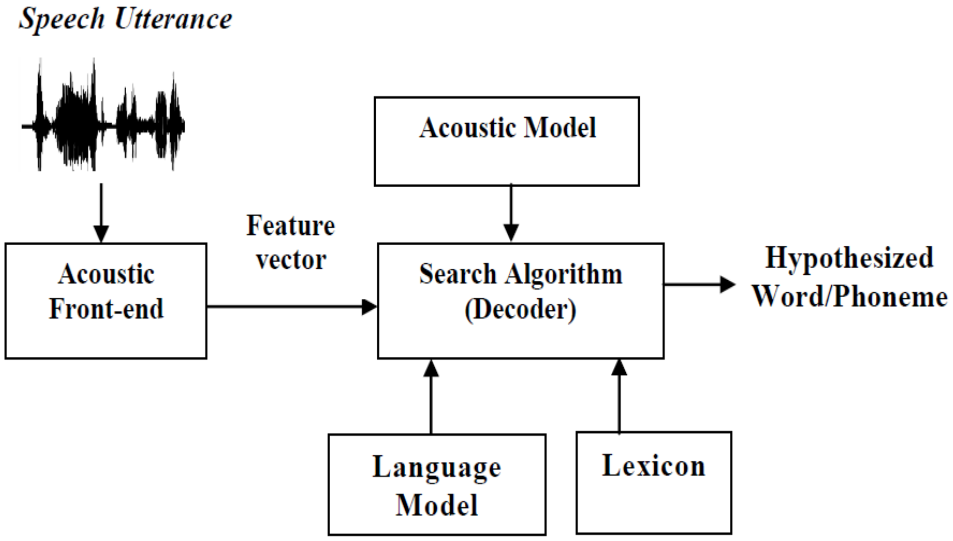


Figure 2.2 – ASR architecture (Karpagavalli and Chandra, 2016: 395)

According to this architecture definition, the so-called **acoustic front-end** or module is responsible for the conversion of the speech signal into features, which feed into the recognition process. In more technical terms, the waveform generated by the audio input of the microphone is turned into a series of acoustic vectors generating the process of features extraction. This extraction is made possible in three stages, as explained by Karpagavalli and Chandra:

> *"The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectra temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage (which is not always present) transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer." (2016: 395).*

More specifically, without entering into more detail regarding the feature extraction, the mechanism of speech features selection is usually performed considering the possibility of discriminating between different, though similar sounding speech sounds, the automatic creation of acoustic models for these sounds, as well as the necessity of exhibiting *"statistics which are largely invariant across speakers and speaking environment"* (*ibid*). As explained by Karpagavalli and Chandra (2016: 395), the likelihood of speech features extraction is defined (in probabilistic terms) as an acoustic model and the conversion process is regulated by the language model.

As far as the **acoustic model** is concerned, according to Karpagavalli and Chandra (2016), it represents *"one of the most important knowledge sources for automatic speech recognition system"* (p.395) and it is responsbile for the representation of the *"acoustic features for phonetic units to be recognized"*. This was also confirmed in the works by Lewis (2015) and by Ghai and Singh (2012). For the building up of the acoustic model, the selection of the basic modeling units is necessary. As seen in §2.2.2 above, the Hidden Markov Model (HMM) is one of the most commonly used statistical models to build acoustic models. But recently, other acoustic models have started to include different models such as the neural networks, as already described above.

According to Ghai and Singh (2012), the **language model** is the other second important element in the ASR architecture. To better understand it, it is possibile to use the decription by Karpagavalli and Chandra who define it as:

> *"A collection of constraints on the sequence of words acceptable in a given language. These constraints can be represented, for example, by the rules of a generative grammar or simply by statistics on each word pair estimated on a training corpus"* (2016; 395).

However, in addition to offering this simple definition of the language model based on grammar rules and statistics, the authors had also the merit of identifying the main function of a language model in the ASR technology, that is to say the feeding of context into the speech recognition process. In fact, for the first time, they defined the language model not only on the basis of the grammar rules or statistics as previous authors did, but they also highlighted the importance of the fact that humans generally add some context information to sounds and words in order to properly recognize the speech units. Therefore, the feeding of context into ASR systems represents one of the main functions to be considered in a language model. The language model in fact helps in indicating what are the valid words in the language and in what sequence they can occur. But how this happens is regulated by an argorithm and this function is played by the decoder according to the architecture defined by Karpagavalli and Chandra.

The **ASR decoder** (also called "Search Algorithm") has the function of searching for the most probable sequence of word units according to a probablistic algorithm. In fact, according to Karpagavalli and Chandra: *"in the decoding stage, the task is to find the most likely word sequence W given the observation sequence O, and the acoustic-phonetic-language model"* (ibid: 398).

To complete the description and review of ASR architecture and its functioning, it is important to add that various approaches and types of speech recognition systems have come into existence in the last decades. This evolution can be described, together with Karpagavalli and Chandra (2016) and Ghai and Singh (2012), as incorporating a series of typologies of approaches, but the present study will only review the main ones: Acoustic-Phonetic approach, Pattern recognition approach, Artificial Intelligence Approach (also known as Knowledge based approach), Connectionist Approach, the Deep Learning,

As far as the **acoustic-phonetic approach** is concerned, it was Hemdal and Hughes (1967) who, for the first time, proposed that spoken language includes a fixed number of distinctive phonetic units that can be generally characterized by a set of acoustic properties varying with respect to time, within a speech signal. According to this approach, the message bearing the units of speech incorporate a series of acoustic properties such as nasality, frication, voiced-unvoiced classification and continuous features such as formant locations, ratio of high and low frequencies. However, as commented in Ghai and Singh (2012: 42), *"for commercial applications, this approach hasn't provided a viable platform"*.

Probably the most important approach in ASR is the so-called **pattern recognition approach**. As already seen in §2.2.2 above on the definition of ASR technology, this approach is considered as the most relevant one in ASR technological evolution. Probably, it was Itakura (1975) who for the first time proposed this approach for the acceptance among researchers. As commented in Ghai and Singh, *"this approach has become the predominant method for speech recognition in the last six decades*" (2012: 42). According to this approach, the main steps are the pattern training and pattern comparison based on a well formulated mathematical framework, which is the distinctive feature, acording to the two scholars. More specifically, the speech pattern representation may take the form of a speech template or a stochastic model: the former leads to a **template-based approach**, while the latter leads to a **stochastic approach**. Within the template-based approach, as described by Ghai and Singh, *"a collection of prototypical speech patterns are stored as reference patterns which represents the dictionary of candidate words"* (2012: 43). In particular, *"an unknown spoken utterance is matched with each of these reference templates and a category of the best matching pattern is selected"* (Ibid.). The advantage of this mechanism is that errors connected with small acoustic units such as phonemes can be avoided. In fact, as argued by Ghai and Singh, *"usually template for each word is constructed" (ibid)*. In other words, every word *"must have its own full reference template" (ibid).* Yet this kind of template preparation and matching can become *"prohibitively expensive or impractical as vocabulary size increases"*, as highlighted again by Ghai and Singh (Ibid.). For this purpose, De Watcher et al. (2007) proposed to resolve this problem by discarding the information about time dependencies and over-generalisation, and by applying a template-based continuous speech recognition with DTW. Finally, in the other model of the pattern recognition model, that is to say the stochastic approach, the functioning of ASR is based on the use of probabilistic models. In this way, uncertain or incomplete information (e.g., confusable acoustic units or homophones) can be dealt with. Together with Ghai and Singh, it is possible to maintain that HMM-based stochastic modelling is *"more general and possesses firmer mathematical foundation in comparison to template-based approach"* (2012: 43).

A more recent, innovative approach is probably the so-called **Artificial Intelligence approach**, denominated by Ghai and Singh (2012) as **Knowledge-Based approach**. This approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In particular, to use the definition offered by Ghai and Singh:

> *"This approach focuses on to mechanize the speech recognition process according to the way a person applies intelligence in visualizing, analysing, and characterizing speech based on a set of measured acoustic features". (2012: 43).*

According to these scholars, both the acoustic phonetic and the template-based approach *"failed at their own to explore considerable insight into human speech processing"* (Ibid.). On the other hand, with the new Artificial Intelligence) approach, knowledge helps the algorithm to *"perform better and also in the selection of a suitable input representation, the definition of units of speech and the design of the recognition algorithm itself"* (Ibid.).

At this point, it is therefore important to understand what "knowledge" means according to the Artificial Intelligence approach: i.e., additional information or input to be entered into the system in the form of a database. For example, Samoulian (1994) presented a data-driven methodology where the knowledge about the structure and characteristics of the speech signal is acquired explicitly from a database. On the other hand, Tripathy et al. 2008 offered a knowledge-based approach using a set of data with spoken English vowels for their classification and recognition. But these are just a few examples of knowledge which can become input within the context of speech recognition. In general, Karpagavalli and Chandra (2016) underline that the main methodologies that contributed to a significant change are the deterministic pattern matching based on dynamic time warping (DTW), and the stochastic pattern matching employing hidden Markov models (HMMs). As a matter of fact, in state-of-the-art systems, *"HMM-based pattern matching is preferred instead of DTW due to better generalization properties and lower memory requirements"*, according to Karpagavalli and Chandra (2016: 399).

Knowledge also plays an important role in the so-called **Connectionist approach**. In fact, this approach focuses on the representation of knowledge and on the integration of knowledge sources. Within this approach, probably the youngest development in speech recognition, "*knowledge or constraints are distributed across many simple computing units rather than encoded in individual units, rules, or procedures"* (Ghai and Singh, 2012: 43). More specifically, it is called "connectionist approach" as knowledge is identified in the *"connections and interactions between linked processing elements"* (ibid). The processing phase of data computation is carried out by a sort of networks of these units, in a similar way to what happens in the human nervous system. The mechanism of the connectionist learning modality is aimed at optimizing that network of processing elements.

Finally, it is important to mention the **Deep Learning** approach, partially described in §2.2.2. This approach in fact represents an innovative architectural system of machine learning, and it has certainly become a mainstream technology for speech recognition. This approach can be further subdivided into two categories, i.e. generative deep architectures, and discriminative deep architectures. Under this approach, it is possible to find a third typology of architectural approach consisting in the so-called hybrid deep architectures. According to Yu and Deng (2015), and to Hinton et al. (2012), under the hybrid approach, the main deep learning architecture is discriminative, but it is assisted with the outcomes of generative architectures. In the hybrid configuration, it is thus possible to maintain that *"the generative component is mostly exploited to help with discrimination as the final goal of the hybrid architecture"* (Karpagavalli and Chandra, 2016: 399).

After having presented the different components and architecture of ASR technology, a review of some of the marketed ASR technologies is carried out in order to complete the state of the art.

2.2.3.2. ASR Technology review

As already described in previous section, ASR systems can convert a speech signal from a speaker or more speakers into a sequence of words, either for text-based communication purposes or for device controlling. The purpose of evaluating ASR systems is that of quality-checking the *"performance of the systems in order to* *measure their usefulness and assess the remaining difficulties, especially when comparing different ASR systems"* (Errattahi et al., 2018: 32). When reviewing ASR technology, first of all, it should be underlined that speech is one of the most difficult genres in computational linguistics (Goldwater et al., 2010: 181). Prosody, vocabulary and disfluency factors do in fact increase error rates. Although it was ascertained by many scholars that ASR has significantly improved in the last years (see for example Errattahi et al., 2016: 1), accuracy must be further investigated to verify if this technology can be implemented at institutional levels. In particular, human factors or other speaker-dependent variables such as language proficiency, disfluency and canonical or non-canonical pronunciation can alter the final output (Goldwater et al., 2010: 181). Other external factors such as background noise may also influence results. Most advanced software solutions available in the market of ASR technology can now better cope with these factors, and ASR technology based on Deep Learning technologies (*i.e.*, Deep Neural Networks or DNN) are now capable of providing *"transcription with an acceptable level of performance"* (Errattahi et al., 2016: 1). This technological innovation certainly facilitates the integration of ASR technology into many institutional applications such as, for example, in meeting and lectures transcription, speech translation and so on.

Apart from the typical, widely-recognized features of ASR (*e.g.*, speaker-independence, an easy-to-use interface, multilingual acoustic model), it is important to underline that ASR systems have also to comply with the **Large Vocabulary Continuous Speech Recognition (LVCSR)** requisite, which today represents a *"particular challenge to ASR technology developers"* (Errattahi et al., 2016: 1). According to this requisite, the ASR technology must include a large vocabulary for the source language (at least 65,000 words), as well as providing for the signal extraction and processing mechanism developed in a continuous manner (as described more in detail in Saon and Chien, 2012: 1-2).

Starting from these preliminary considerations, the present study's ASR technology review led to examination and testing of some of the best-in-class technologies available on the market (see below), by also trying to meet the requisite of convenience in terms of costs and time/efforts required in implementing this technology in an ordinary-use computer or workstation. More specifically, the review pointed out that currently marketed ASR technologies to be effective must meet the main requisites reported in Table 2.1 below.

Requisites	Description
Speaker-independence	The ASR solution recognizes the voice of different speakers
Easy to use interface	The interface is intuitive also for non-IT experts
Minimal computer requirements	The computer or workstation requirements are of average market level
Multilingual acoustic model	The ASR solution recognizes the two languages of this study: English and Italian (in addition to other 32 languages)
LVCSR	The Large Vocabulary Continuous Speech Recognition requisite is met (see above)
Augmented Terminology	Possibility of uploading domain-specific vocabulary or terms
Cloud-Based	The ASR solution is also usable via Internet and cloud-based
Trainable	The ASR solution can be "trained": it learns from previous processing workflows to obtain better accuracy

Table 2.1 – Requisites for the selection of this study's ASR solution

In the initial phase of the ASR technology review, several solutions were considered, including *VoxSigma* by Vocapia Research[1], Microsoft Skype Translator[2], Google Speech Recognition (via YouTube and Descript)[3] and *Dragon Naturally*

[1] Vocapia Research's VoxSigma official website: http://www.voxsigma.com/speech-recognition-software.html

[2] Microsoft (Skype) Translator: https://www.skype.com/en/features/skype-translator/

[3] Google's Cloud Speech to Text official website: https://cloud.google.com/speech-to-text

Speaking[4] powered by Nuance. After a preliminary phase of testing and usage of all four software solutions, the selection led to the exclusion of *Microsoft Skype Translator* and *Dragon Naturally Speaking* powered by Nuance. In fact, the former was excluded mainly because of the high cost/fee associated to its usage (a free version is available for a few files processing only), while the latter was excluded because of its speaker-dependent ASR engine, not allowing for automatically transcribing speeches from the voice of different speakers. The selection was thus oriented towards *VoxSigma* application developed by Vocapia Research, and Google Speech Recognition engine (via *YouTube* and *Descript* applications) provided that these ASR solutions respond to the requisites described in Table 1 (above) and for the reasons explained below. Additionally, it should be clarified that this study does not intend to promote any particular software or ASR solution as there may be other solutions in the market which could respond to the same criteria above and be used for the same purposes and applications. Therefore, this selection should not be considered as exhaustive. A detailed description is offered in the sections below only for the solutions that passed the preliminary testing phase.

2.2.3.2.1. VoxSigma by Vocapia Research

As far as Vocapia Research's *VoxSigma* solution is concerned (Vocapia Research, 2020), in addition to the ordinary, taken-for-granted ASR features (see Table 2.1 above), *VoxSigma* includes adaptive features allowing the transcription of noisy or disturbed audio files like speeches with background music or applauses, eliminating any disturbance or interference in the limited portions of the files where this noise is present. Additionally, it should be highlighted that, although *VoxSigma* can be used/tested for free for a limited volume of words/minutes, it can offer users with its highest potential when subscribing to the paid service (though at a reduced cost). More interesting is the fact that this solution provides for an Augmented Terminology feature, that is to say it offers the possibility of creating/adding domain-specific databases for adapting the subtitle transcriptions to specific domains. And this feature is of particular interest for this study as the "significance" of terminological resources is assessed and evaluated.

[4] Dragon Naturally Speaking by Nuance: https://www.dragon-naturally-speaking.com/

The software can be used as a *SaaS* ("Software as a Service") solution via the YobiYoba platform[5] and, in the initial phase of the testing (see Chapter 4), it is intended to be used "*as it is*" (*i.e.*, without adding any specific terminology). As described in the official software website, *VoxSigma* software suite offers large vocabulary multilingual capabilities with state-of-the-art accuracy (Vocapia Research, 2020). It is specifically designed for professional users, needing to transcribe large quantities of audio and video documents such as broadcast data, either in batch mode or in real-time (ibid).

Like in other ASR technologies, the complete voice-to-text process is completed in three steps. Firstly, the software identifies the audio segments containing speech, and then it recognizes the language being spoken (if it is not known a priori or set by default), and, finally, it converts the speech segments to text and time-codes. The fully annotated XML document obtained (including speech and no-speech segments, speaker labels, words with time codes, high quality confidence scores, and punctuations, if required) can be converted into plain text (as in the present study).

Among the variety of features/services offered by *VoxSigma*, the following ones are to be mentioned as they may prove to be useful for the next phase of this study. Considered the potential usage at an institutional level, in a real-time, simultaneous modality or in a asynchronous sequence, it is important to report the following features (Vocapia Research, 2020):

> ***"Protocol:*** *REST API over HTTPS; POST, GET and PUT HTTP methods are accepted; both URI encoded requests and MIME multi-part requests are supported; three submission modes: file, streaming, and real-time.*
>
> ***Service Availability:*** *the service is available 24/7/365 with failover servers and geographic redundancy.*
>
> ***Supported audio file formats:*** *AAC, AIFF, ASF, FLAC, MS-Wave, MPEG, Ogg/Vorbis, Nist Sphere, Sun AU*

[5] YobiYoba: https://www.yobiyoba.com/en/

Typologies of audio sources or communication: *via telephone or broadcast quality, real-time production.*

Communication or audio duration per request: *up to few hours (depending on the coding rate).*

Functions: *automatic language identification, audio and speaker segmentation, speech-to-text conversion, and speech-text alignment.*

Generated output: *XML files with speaker diarization, language identification tags, word transcription, punctuation, confidence measures, numerical entities and other specific entities.*

Special features: *on-the-fly language model adaptation, daily updates of language models for broadcast data*

Transcription of speeches: *VoxSigma is currently used by several governmental organizations to provide easy access to video content by generating time-coded searchable XML documents." (Vocapia Research, 2020)*

2.2.3.2.2 Google Speech Recognition engine

With regard to Google Speech Recognition (GSR) solution (available via *YouTube* or *Descript application*), it is necessary to observe that this solution is very popular among users (also for non-IT experts), both at academic and institutional contexts. Apart from the common ASR features required for a sophisticated ASR engine, GSR technology (as reported in the webpage of Google Cloud Speech to Text, 2020) via *YouTube* or *Descript*[6] can offer immediate and easy-to-use transcription functionality directly on the platform website (Descript, 2020), even for audio/video files not published on it yet. In fact, the user can simply upload a file on the platform, and he/she can then carry out an automatic transcription of the file without having to pay any cost or fee (but a free registration is required both for YouTube and for Descript users). Yet it should also be mentioned that, like Vocapia Research's solution, GSR technology (as mentioned in Google Cloud Speech to Text, 2020) allows for a higher level of functionality when subscribing to the paid service (via its Cloud API). This would

[6] Descript API and Web service: https://www.descript.com/

offer for adaptive features like the possibility of uploading specific terminological resources (the Augmented Terminology requisite seen above).

Additionally, as reported on the official website of Google Cloud Speech to Text website (2020), among the various expanded functionalities offered by it (through the *YouTube* platform or as Cloud API), it is here important to underline that:

> *"Speech-to-text conversion is powered by machine learning, and it is available for short-form or long-form audio files. Its powerful Speech Recognition technology allows converting audio to text by applying powerful neural network models in an easy-to-use API. This ASR system can recognize up to 120 languages and variants to support a global user base. More in detail, being powered by machine learning, this ASR system applies the most advanced deep-learning neural network algorithms to audio for speech recognition". (Google Cloud Speech to Text, 2020)*

Not less important is the fact that the Cloud Speech-to-Text accuracy can improve over time as Google improves the internal speech recognition technology used by Google products (the Trainable requisite seen in Table 2.1 above, which is also described for deep learning neural networks system in §2.2.2). In the paid version of GSR engine via Descript interface, the "training" functionality is also available for final users. Additionally, like *VoxSigma*, GSR engine *"can identify what language is spoken in the utterance (limited to four languages per time), which allows returning text transcription in real time for short-form or long-form audio materials"* (see Google Cloud Speech to Text, 2020). This may be of particular interest for institutional communications as in the setting considered for this study. As it is possible to read on the official website, this technology can also return automatic transcriptions in a file format, and *"it automatically transcribes proper nouns"* (Google Cloud Speech to Text, 2020). For example, it is said to be tailored to *"work well with real-life speech and it can accurately transcribe proper nouns (e.g., "Sundar Pichai") and appropriately format language (e.g., dates, phones numbers)"* (ibid). Again, like in *VoxSigma*, speech recognition can be customized to a specific context by providing a set of words and phrases that are likely to be spoken (paid service only), thus responding to the Augmented Terminology requisite seen in Table 1 above. This is especially useful for adding custom words and names to the vocabulary, or specific terminological resources. Multiple audio encodings are supported, including FLAC, AMR, PCMU, and Linear-16 (ibid).

To conclude, it is here interesting to highlight the following features which could be useful for this study and for an institutional communication setting (see Google Cloud to Text, 2020):

- ***"Noise Robustness:*** *it handles noisy audio from many environments without requiring additional noise cancellation.*
- ***Automatic Punctuation:*** *it can accurately punctuates transcriptions (e.g., commas, question marks, and periods) with machine learning (in VoxSigma, only end-of-sentence punctuation is reported).*
- ***Model Selection:*** *it is possible to choose from a selection of four pre-built communication models: default, voice commands and search, phone calls, and video transcription." (Google Cloud Speech To Text, 2020).*

2.2.3.3. Studies on previous ASR projects

2.2.3.3.1. Verbmobil

The first project conducted within the European Union on the use of Automatic Speech Recognition technology is the 8-year *Verbmobil* project, which was substantially described in Wahlster (2000). The project started in 1992 and ended in 2000, with the collaboration of 31 different partners in three continents. The final outcome was the same-name software system that provided mobile phone users with simultaneous dialogues interpretation services for restricted topics. As described in Wahlster (1993) and Kay et al. (1994), this project was promoted by the German Government (Federal Ministry for Education and Research - BMBF) and controlled by the German Aerospace Research Establishment (DLR), Berlin, with the collaboration of the Artificial Intelligence Research Centre (DFKI GmbH) based in Saarbrucken. The aim of the project was to develop and investigate on the potential application of Speech-to-Speech technology for mobile conversations in three domains (mainly business-oriented), and with a focus on three main languages: English, German and Japanese. At that time, the German Government's goal was to spread the usage of the German language within the European Union and internationally in the business industry.

The use of Verbmobil software system was completely hands-free, since it did not require users to press any push-to-talk button. Since the Verbmobil speech translation server could be accessed by any GSM mobile telephones, the system could be used anywhere and anytime. From a technical point of view, a significant campaign of data collection was performed during the Verbmobil project to further improve and train the statistical model incorporated into the system on the basis of corpora of spontaneous speech. A distinguishing feature of Verbmobil was probably its engine (statistical) and the fact that it represented the "*first spoken-dialog interpretation system to use prosodic information*" (Wahlster, 2001: 1484). In its final version, Verbmobil system also included a multiple translation engine which covered a wide spectrum of translation methods. To better identify the main features and properties of Verbmobil system, all its features are summed-up in Table 2.2 below (information collected from Wahlster, Ed., 2000).

Feature	**Description**
Speaker-independence	The system can recognize any speaker's voice.
Bidirectionality	Conversation can be bidirectional (from speaker 1 to speaker 2 and vice versa).
GSM-based	The system is based on GSM mobile technology.
3 languages	The system can recognize and translate English, German and Japanese.
Specific vocabulary	The vocabulary is business-specific and includes over 10,000 words.

Table 2.2 – Main features of Verbmobil software (Wahlster, 2009)

Certainly, this project is worth mentioning as it was the first attempt to combine a Machine Translation system with Automatic Speech Recognition technology. In all previous studies, these technologies were kept separated and the outcomes produced by their combination were not examined on an aggregate basis. Apart from that, this project was also the first study in which the disfluency elements of speech were considered and treated in the processing and analysis of data. As explained in Chapters 3 and 4, the disfluency elements of speech are in fact an essential feature to be analysed for an evaluation of accuracy. Finally, even if it was carried out in a limited way, this project offered, for the first time, considerations on the importance of vocabulary (more specifically, the business vocabulary) for the generation of ASR output. In addition to the weaknesses of this project connected with the limited RAM capacity of the device implemented, and the reduced span of its vocabulary, it should be concluded that, although the Verbmobil project incorporated a domain-specific vocabulary, its evaluation strategy did not include the impact of that vocabulary in the analysis. Furthermore, the purposes of accessibility were not included among the objectives of the project.

2.2.3.3.2. TC-STAR

The TC-STAR project focused on the automatic translation of European Parliamentary speeches, as described in Vilar et al. (2005) and in Fügen et al. (2006). The TC-STAR project, financed by the European Commission within the Sixth Framework Programme and with the coordination by *Istituto Trentino di Cultura* (ITC), was a 36-month initiative started in April 2004. TC-STAR was envisaged as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation (SST), as defined in the official project website (European Commission, 2004-2007)[7].

Similarly to Verbmobil, TC-STAR included a combined usage of Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech (TTS) (speech synthesis). However, unlike Verbmobil, the domain was much wider (not only focused on business-oriented communication) and it included a selection of speech domains into three languages (European English, European Spanish and

[7] European Commission (2004-2007). TC-STAR. Technology and Corpora for Speech to Speech Translation. http://www.tcstar.org/

Mandarin Chinese). At the early stage of the project, TC-STAR was focused on the translation of specific speeches delivered during the European Parliament Plenary Sessions (EPPS). This made TC-STAR the first European project on spoken language translation working on a real-life task. In particular, two translation directions were considered: from English to Spanish and from Spanish to English. In a subsequent phase, the project analysed the translation of broadcast news for the English to Mandarin Chinese combination. The software used was developed within the partner universities and research centres[8] and the ASR results were translated by using a MT solution: i.e., *Systran*. By comparing the input speeches with the MT output, the efficacy of the solution was evaluated. In fact, as indicated by Mostefa et al.:

> *"To evaluate the performance of a complete speech-to-speech translation system, we need to compare the source speech used as input to the translated output speech in the target language. The proposed methodology enables to measure the fluency and the adequacy of the translated output".* (*Mostefa et al. (2006: 1).*

This methodology was used, for example, for the evaluation of English-to-Spanish direction. For this part of the project (the most institutional one and thus the most relevant for this study), the data consisted of audio recordings in English of the European Parliament Plenary Sessions (EPPS), where the focus was on Members of Parliament speaking in English. In particular, the evaluation data were made of 20 segments of around 3 minutes each. Therefore, in total, the evaluation set of data was composed of 1 hour of speech and around 8,000 running English words.

The most important aspect in TC-STAR was probably the fact that the three main components (ASR, SLT and TTS modules) were trained on data including training corpora built from the EPPS (European Parliament Plenary Sessions) recordings. For each audio sample in English an ASR output was produced, then the

[8] The Consortium included: Istituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica (ITC), Rheinisch-Westfaelische Technische Hochschule Aachen (RWTH-AACHEN), Centre National de la Recherche Scientifique (CNRS-LIMSI), Universitat Politècnica de Catalunya (UPC), Universitaet Karlsruhe (TH) (UKA), IBM Deutschland Entwicklung GmbH (IBM), Siemens Aktiengesellschaft (SIEMENS), Nokia Corporation (NOKIA), Sony International (Europe) GmbH (SONY). Source: ELRA, 2015: http://www.elra.info/en/projects/archived-projects/tc-star/

ASR output was automatically translated into Spanish and, finally, the SLT output was synthesized in Spanish by the TTS module using the alignment between SLT and ASR to get the prosodic features from the source language.

As far as the evaluation process was concerned, the project included different evaluation stages at the end of each single study or phase (throughout its duration). The main tools used were the metrics generally implemented in machine translation studies (see §2.4. on Machine Translation), as well as evaluations based on questionnaires. For each phase, using a specific protocol, the evaluation comprised three steps: 1. first, a questionnaire was established for each English sample; and then the sample was translated into Spanish; 2. then, evaluators assessed the Spanish translated samples according to the evaluation protocol; 3. finally, the subjective evaluations results (answers given by judges) were checked by a single person. The study evaluation process also tried to compare the TC-STAR system's outputs with the output of professional interpreters, and to do that correctly, judges involved in the process were not informed about the presence of audio data from interpreters or when it was produced by the TC-STAR system.

In general, this project's conclusions offered various hints and suggestions for future work. The best tool offered is probably the evaluation methodology, including a combination of MT metrics and subjective questionnaires, as well as a comparison with professional interpreters. Another important aspect of this project is represented by the wide selection of EU-related corpora and data collected under it. In addition to the European Parliament speeches, the study incorporated recordings from Europe by Satellite channel, texts from EU Translation Service, transcriptions of EP speeches, Spanish Parliament (*Cortes*) speeches (to expand the Spanish database), as well as EU Parliament's Final Text Editions (FTEs). As a conclusion of this section, the main features and innovations introduced by TC-STAR project are summed up in **Table 2.3** below.

Feature	**Description**
Computer-based software	The system is based on computer technology

Institutional scenario	The system is applied to an institutional context (i.e., the European Plenary Sessions)
SMT	The Machine Translation used has a statistical engine
3 languages	English, Spanish and Chinese (Mandarin)
Comparison with other SMT solutions	The system's output is compared to a market SMT solution: *Systran*
Comparisons with Interpreter's performances	The system's output is aligned and compared to a collection of interpreters' renditions
Definition of a pipeline	The project develops in 3 steps: ASR > SLT >TTS
Quantitative and qualitative evaluation methodology	Questionnaires are used for a qualitative analysis, while other metrics (e.g. BLEU[9] index, WER rate, etc.) are implemented for a quantitative analysis.

Table 2.3 - Main features of the TC-STAR project.

In general terms, it is possible to maintain that the TC-STAR project included features similar to this study with respect to the institutional focus (the analysis of EP speeches) and the usage of a pipeline which is partially common to the present study (as illustrated in Chapter 3). In addition, one could say that the TC-STAR project offered interesting hints and outcomes to the present study also in relation to the methodology adopted and the metrics used. In fact, the project proposed a combined methodology of quantitative metrics (namely, the BLEU[10] index and the WER rate) and qualitative tools, *i.e.*, the usage of questionnaires. However, in this respect, it should be pointed out that the limited length of the segments examined (3 minutes each) did not make it possible to examine the entire context of the speech material and the overall criticalities of ASR (for example, the Segmentation phenomenon described in Chapter 4, §4.7). Furthermore, the implications of ASR errors on MT output are not

[9] BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. This metric is considered as the de facto standard automatic evaluation metric in machine translation (Song et al., 2013).

[10] For further information on this algorithm, consult §4.5 and §4.9.

considered at all and the usage of questionnaires may be limited in its scope by issues of subjectivity. More specifically, for the quantitative part of the analysis, the BLEU metric cannot be considered as an effective measure of accuracy, as commented in Song et al. (2013): *"While BLEU is undeniably useful, it has a number of limitations. Although it works well for large documents and multiple references, it is unreliable at the sentence or sub-sentence levels, and with a single reference."* Other weaknesses of this project are connected with the implementation of an SMT technology (offering lower quality output if compared to current state-of-the-art NMT systems), as well as with the fact that the impact of terminology are not examined. Finally, given the accessibility scope of the present study, it is necessary to underline that the TC-STAR did not focus its attention on accessibility implications and on the benefits of ASR implementation for accessibility purposes.

2.2.3.3.3. EU-BRIDGE

This institutionally-based project was another European umbrella research project aimed at developing different research activities providing innovative speech translation technology. Running from 1st February 2012 to 31st January 2015 under the supervision of CORDIS (Community Research and Development Information Service), EU-BRIDGE (*"EU-Bridges Across the Language Divide"*) was funded by the European Union under the Seventh Framework Programme (FP7)[11]. The partners of EU-BRIDGE included the Aachen University (RWTH), the University of Edinburgh (UEDIN), the Karlsruhe Institute of Technology (KIT), and the Fondazione Bruno Kessler (FBK), which participated into the project's large-scale evaluation campaigns.

Like the present study, the initiative stemmed from the idea that the current production of multilingual content now far outpaces the EU institutions' ability to have it translated by humans, hence the necessity of turning to automatic methods to cope with this need. Specifically, EU-BRIDGE focused on 4 use cases: *i)* interlingual subtitles translation for TV broadcasts; *ii)* the translation of University lectures; *iii)* the

[11] European Commission Community Research and Development Information Service (CORDIS), "Seventh Framework Programme (FP7)". http://cordis.europa.eu/fp7/.

translation of speeches at the European Parliament; *iv)* Unified Communications Translation.

For the purposes of this study, only the part of the EU-BRIDGE project focusing on the automatic translation of the European Parliament's speeches will be reviewed. Within this specific use case, EU-BRIDGE applied language technologies to the interpreting workflows at the European Parliament. In practical terms, the EU Parliament's interpreters were supported in their preparation for meetings and in their interpreting service by providing terminology resources and automatic translations which consisted in a set of tools including ASR and MT systems. This set of tools was realized and supplied as a Web application in which interpreters could find their preparation documents analysed by the system (and machine translated). In particular, with the Web app, special-domain terminology and named entities were provided as automatically identified and appropriate translations were also suggested from a variety of sources, such as online sources or the translation memory of the European Commission.

One of the most interesting initiatives within the EU-BRIDGE project was focused on Automatic Speech Translation (AST) between the English-French and the German-English language pairs. In this specific subproject, the developed system could generate automatic translation of European Parliament speeches into these languages. As reported in Freitag et al. (2013) during the 2013 International Workshop on Spoken Language Translation (IWSLT)[12], as far as the audio and written corpora are concerned, this subproject incorporated a large amount of publicly available monolingual and parallel training data (WIT, Europarl, Multi-UN, the English and French Gigaword corpora) to improve the output quality. However, as indicated in Freitag et al. (2013), this initiative had a strong focus on translation of spoken language, while little attention was given to the phase of Automatic Speech Recognition and to its evaluation. Additionally, as described in Matusov et al. (2006), it provided a combination of different MT engines from the partner Universities (RWTH, UEDIN, KIT, and FBK) involved. This systems combination was used to produce *"consensus translations"*, that is to say the possibility of computing a consensus translation from the outputs of multiple machine translation (MT) systems. According to this method, as explained by Matusov et al., *"the outputs are combined*

[12] International Workshop on Spoken Language Translation 2013, http://www.iwslt2013.org.

and a possibly new translation hypothesis can be generated" (Matusov et al., 2006: 33). Additionally, as commented by Matusov et al. (2008: 1222), *"consensus translations can be better in terms of translation quality than any of the individual hypotheses"*. And the conclusion by these scholars is also backed by others, such as Freitag et al., who commented that:

> *"While each of the individual engines provides performance that is state-of-the-art for single systems, the results suggest that system combination techniques are still a fertile approach to benefit from diversity in collaborative efforts and thus progress towards even better quality"*. (2013: 6)

Considered the fact that EU-BRIDGE project results were published in different works relating to each specific use case (and that it is difficult to have a general overview of the final results), here it is sufficient to say that, by joining the outputs of the partners' different individual machine translation engines via a system combination framework, the final output was significantly improved in terms of translation performance (up to +1.4 BLEU and -2.8 TER[13]), as reported in Freitag et al. (2013: 5). To conclude with the description of the EU-BRIDGE project, the main features are now summed up in Table 2.4 below.

Feature	**Description**
Institutional scenario	The project is defined and implemented within European institutions and universities
Multi-project structure	4 use cases
Web application	The set of tools is offered to interpreters as a Web interface to be used during their work
Support of Translation Memory	It is based on supportive translation memory, including EU corpora material

[13] TER: Translation Error Rate is a method used by Machine Translation specialists to determine the amount of Post-Editing required for machine translation jobs. The automatic metric measures the number of actions required to edit a translated segment in line with one of the reference translations. It's quick to use, language independent and corresponds with post-editing effort (KantanMT Blog, 2015).

Two language pairs	English-German and English-French
Multiple MT system	The system provides for the application of multiple MT systems for a better output: “consensus translation” architecture

Table 2.4 - Main features of the EU-BRIDGE project.

Apart from the institutional scenario offered within this initiative (which is similar to the present study’s scenario and context), the most important contribution of EU-BRIDGE to the research on ASR is probably the usage of a “consensus translation” based architecture for the part of the process involving the implementation of a MT system. However, in this respect, it should be underlined that, again, the MT system adopted here is an SMT technology which may today appear as not sufficiently effective if compared to NMT (to be considered as the state-of-the-art technology for this study). The outcomes generated from this project are certainly useful for the evaluation of an SMT architecture (for example, the usage of the TER index in addition to the BLEU metric), if also accompanied with the usage of multiple institutional Translation Memories and interpreting parallel corpora, but they may not be considered relevant as little attention was dedicated to ASR analysis. As a matter of fact, ASR technology was mainly used as a tool for the interpreting work into the booth and the relevant output was not examined in relation to MT, neglecting the evaluation of ASR output. So, to conclude, for the purposes of the present study, the scope of the EU-BRIDGE initiative is limited and it can only be considered for the Machine Translation component of the pipeline, as well as for the phenomena and criticalities connected with the spoken language translation. Finally, another interesting hint for the present study is probably the relevance of internal terminology and vocabulary in the implementation of the solution for institutional interpreters at the EP, tough little evidence was offered with respect to the impact of terminology and domain-specific vocabulary for the improvement of accuracy in the ASR phase of the process. In fact, the analysis of results was only focused on SMT accuracy without offering considerations for the ASR output.

2.2.3.3.4. DARPA-GALE

Outside the EU context, a significant scientific project is represented by the DARPA-GALE initiative promoted by the U.S. Government and Defense Department. In particular, DARPA (Defense Advanced Research Projects Agency) promoted two strategic initiatives concerning translation and automatic speech translation: namely, the BOLT (Broad Operational Language Translation)[14] and the GALE (Global Autonomous Language Exploitation)[15] initiatives. The first initiative was a military and defence-oriented project aimed at developing genre-independent, machine translation and information retrieval systems, not pertaining to speech recognition. Therefore, in this section, our attention is only focused on the GALE project directed by SRI International, which is an independent, private US Research Centre.

Started in 2005, the DARPA-GALE programme aimed at producing a system capable of automatically taking multilingual newscasts and text documents, and to make their information available to human queries: i.e., offering the possibility of searching for specific phrases or terms (the Distillation feature described below). In particular, GALE coped with three major technical challenges, as commented in Anderson (2006): Automatic Speech Recognition (to process audio data), Machine Translation (to translate non-English data) and, as an exclusive component of this project, Distillation (to extract the most useful pieces of information related to a given query). Previous projects or systems had carried out all these steps as separate or sequential processes; on the contrary, DARPA-GALE, to quote Olive et al. (2011: X): *"involves use of a distinctly new approach, one by which researchers have sought to create systems able to execute these processes simultaneously"*. More specifically, DARPA-GALE provided a 3-component architecture where the first module (ASR) handles the transcription of spoken languages into text; the second one (MT) provides for a translation process that can convert foreign text into English, and the third is a "distillation" engine that can answer questions and summarize information coming from the other two modules. As highlighted by Olive et al. (Ibid.: X-XI), the software solution (created in collaboration with IBM and BBN Technologies) implemented a combined statistical and linguistics translation model for MT, an integrated distillation

[14] DARPA. Broad Operational Language Translation (BOLT). https://www.darpa.mil/program/broad-operational-language-translation

[15] DARPA. Global Autonomous Language Exploitation (GALE): http://www.speech.sri.com/projects/GALE/

feature to extract relevant information, and it was based on neural networks (NNs) for the acoustic model (ASR). In addition, the software solution was ready to be used and distributed on any computer as a “plug-&-play” solution, including software to facilitate integration and optimization. This complete ASR-MT system offered for the first time the possibility of combining ASR with MT into three key languages: English, Chinese and Arabic. Additionally, DARPA-GALE was important insofar as it highlighted the necessity of improving vocabulary accuracy. In fact, as claimed by Kumar et al. in discussing the project’s results:

> *“While each of these component technologies have continued to improve in performance, each is data-driven and its performance will degrade when faced with novel vocabulary”. (2015: 229)*

As a matter of fact, even a large vocabulary-based ASR system is incapable of recognizing out-of-vocabulary (OOV) words, and the MT system used cannot translate unseen or misrecognized source words and the TTS module often mispronounces novel words (namely, concepts like proper names and technical terminology). This consideration may prove of particular relevance for the purposes of the present study, which is going to examine the impact of terminology and domain-specific vocabulary in the evaluation of accuracy. To conclude, the main features of DARPA-GALE project are summed up in Table 2.5 below.

Feature	**Description**
Institutional scenario	Military or defence application
Distillation feature	Apart from ASR and MT, the system offers a Distillation feature to extract information
3 languages	English (U.S.), Arabic and Chinese
Relevance of Terminology	The studies highlighted the necessity of improving terminology accuracy in ASR and MT

Acoustic model	For ASR, a neural networks-based acoustic model is implemented.

Table 2.5 – Main features of the DARPA-GALE project.

It is worth underlining that this institutional project is relevant to the present study for being the first institutional project where ASR is effectively combined with an SMT system to produce automatic interlingual subtitles (more specifically, with reference to the automatic translation of TV news), in a sequential and automatic system which is similar to that implemented in this study. Additionally, unlike the previous projects, DARPA-GALE offered, for the first time, a general focus on the impact of terminology, tough little attention was given to its effects on accuracy measurement. Another important aspect of this initiative is represented by the implementation of an ASR system incorporating an acoustic model based on neural networks, which, as described in §2.2.1 and §2.2.2 above, represents an essential requisite for an effective ASR process. However, DARPA-GALE presented a series of weaknesses, mainly consisting in the usage of an SMT system instead of a NMT system and the fact that no sufficient publications on the results are made available, also because of the military and government's nature of the initiative (in fact, most of data were not published).

2.2.3.4. Studies on the combination of ASR with interpreting

A third group of studies on ASR deals with the interaction of Automatic Speech Recognition technology with interpreting and with the interpreters' service/work. In this chapter, only the more recent works carried out in this specific field will be reviewed, including their contributions to the studying of quality assessment.

Before examining these works more in detail, it is important to report about and to describe the new typologies of interpreting service and work within the interpreting industry, which are strictly interconnected and combined with the usage of the most advance technologies. In fact, the wide spreading of Information and Communication Technologies (ICT) across the interpreting industry has recently led to significant changes, including the rise of new forms of interpreting. According to

Fantinuoli, the *"topic of technology is not new in the context of interpreting"* (2018: 1). However, the *"recent advances in interpreting-related technologies are attracting increasing interest from both scholars and practitioners"* (Ibid). Although those studies on new forms of interpreting are generally incorporated in the scientific literature within the disciplinary field of Interpreting Studies, it is here worth considering them to obtain useful hints and conclusions on the combination of ASR and other computer-assisted solutions with the interpreter's work and function.

As already mentioned above, in its long history, interpreting has not been immune to technological innovations. As a matter of fact, according to Fantinuoli, *"it has gone through at least two major technological breakthroughs with disruptive effects on the profession in both cases"* (Fantinuoli, 2018: 2). In line with these considerations, also D'Hayer (2012: 236) confirms that *"new technologies play an innovative role that can no longer be ignored in the world of interpreting"*. More specifically, Pöchhacker regards ASR as a technology *"with considerable potential for changing the way interpreting is practiced"* (2016: 188). In addition to the introduction of wired systems for speech transmission that led to the rise of simultaneous interpreting (SI) in the 1920s, the second, most important technological breakthrough was represented by the arrival of the Internet in 1990s. As already discussed in several works, the arrival of the Web radically changed the interpreters' relation to knowledge and its acquisition. In fact, as in the work of an interpreter the preparation of conference material represents an important phase of his/her work (Gile 2009), the opportunity of finding useful material through the Web had enormous effects on the profession. To use the words of Fantinuoli, *"Internet is the most comprehensive and accessible repository of textual material available in many languages and on many topics"* (2018: 2). For example, recent studies have shown that the Web can be used by interpreters to conduct exploratory research before they receive actual conference material (Chang et al. 2018) or to create specialized corpora for linguistic analyses (Fantinuoli 2017a).

Examining the most recent, advanced technological evolution of the last two decades, it is possible to maintain that interpreting is probably facing and coping with a third breakthrough, which, to quote Fantinuoli, can be defined as a *"technological turn in interpreting"* (2018: 3). This new, important technological transformation is strictly interconnected with the arrival of new interpreting-related technologies:

computer-assisted interpreting (CAI), remote interpreting (RI), and machine interpreting (MI). Without entering into more detail, it is worth mentioning that Remote Interpreting is interconnected with the use of Cloud and Internet technologies, as well as with telephone-based or advanced videoconference equipment for the provisioning of distance or remote interpreting services. Initially, according to D'Hayer, RI was seen as *"a controversial topic still lacking in quality standards and ethics"* (2012). Yet many of today's public services such as the police and the ambulance services use telephone interpreting services every day in great urban areas (D'Hayer, 2012); video conference interpreting and web streaming facilities are also gaining popularity on the private market and also in the justice courts. This has led not only to the definition in the scientific literature of a new dimension and role for the so-called "Community Interpreter" (or "Public Service Interpreter"), but also a series of new studies and projects. For example, the AVIDICUS project (Braun et al., 2016), Assessing videoconference interpreting in the Criminal Justice System (EU Criminal Justice Programme, Project JLS/2008/JPEN/03), followed by the IVY project (Interpreting in Virtual Reality: EU Lifelong Learning Program, Project 511862-2010-LLP-UK-KA-KA3MP) have demonstrated that we need to *"anticipate, prepare, understand and assess interpreting and translation skills within a new virtual dimension"* (D'Hayer, 2012: 237). More recently, the SHIFT project aimed at developing a *"theoretical and methodological framework for the analysis of interpreter-mediated oral discourse in telephone and video-based interpreting"* (2018: 47).

For purposes of this review, the other two typologies of interpreting-related technologies are probably of major relevance: i.e., MI and CAI. Under these two typologies of interpreting, in fact, Automatic Speech Recognition certainly plays and will play an important role in the next future. Computer-assisted interpreting (CAI) can be defined, to quote Fantinuoli, as a *"form of oral translation in which a human interpreter makes use of computer software designed to support and facilitate some aspects of the interpreting task with the goal to increase quality and productivity"* (Fantinuoli 2018a). For example, CAI tools may include instruments used for assisting interpreters in the creation of glossaries to be consulted during their work in the booth, or looking-up tools for rapid terminological searches or even distillation tools to extract useful information from preparatory documents (see Sandrelli and De Manuel Jerez, 2007).

Among these technological solutions and instruments, Automatic Speech Recognition is one of the most impacting technologies for the booth, as this technology makes it possible to produce automatically transcribed source text for consecutive, or even simultaneous interpreting. In particular, in the work of Desmet et al. (2018), the potential impact of CAI tools was analysed regarding a technology which allows for the automatic recognition of numbers in the source speech and which presents them on a screen in the booth. This tool proved to reduce the cognitive load during simultaneous interpreting, as well as to improve quality. Thanks to the evaluation of quality reached by ASR, the present thesis may help in facilitating the adoption of this technology at the booth level. The impact of CAI tools on the interpreter's job is also crucial in Prandi's study, whose preliminary results have indicated that, even if the CAI tools may increase saturation and workload, they may however contribute to increase output quality in terminological terms when used for the consultation of interpreting material (Prandi, 2018). In a recent study, Corpas Pastor and May Fern (2016) have analysed computer-assisted interpreting (CAI) programmes used by professional interpreters to prepare for assignments, to organize terminological data, and to share event-related information among colleagues: they found that the programmes accelerated the process in collecting conference preparation material.

Automatic Speech Recognition (ASR) has also been examined as a technology used to automate the querying system of CAI tools (for example, in Fantinuoli, 2016), which is a totally different application with respect to the usage of ASR for transcribing speeches. Through ASR, the interpreter can in fact look up for terms or identify specific terms without having to type that specific term in the glossary or without having to scroll the glossary list. In addition to this application, thanks to the more recent advances in Artificial Intelligence, especially with the implementation of deep learning and neural networks (as described in §2.2.1 and 2.2.2 above), the quality of ASR has significantly increased (Yu and Deng, 2015). According to Fantinuoli, with the possibility of exploiting systems capable of achieving a 5.5 percent word error rate, the usage of ASR in interpretation is absolutely *"conceivable nowadays"* (Fantinuoli, 2017b: 2). In consecutive interpreting, ASR as a CAI tool may be used to generate an automatic transcription of the spoken word to then sight-translate the speech segment, with obvious advantages in terms of precision and completeness (Fantinuoli, 2017b: 3). In the case of simultaneous interpreting, the ASR technology may be used not only to query the reference materials or glossary previously prepared (as in Prandi, 2018), but also to automatically recognize parts of the speech which are often considered as "problem triggers" in interpretation: namely, numbers, acronyms and proper names. In this specific case, ASR is used as a tool to prepare the conference material, rather than being used as an instrument of automatic speech recognition for transcription. Most importantly, for a successful operation of the ASR solution within the interpreter workstation, ASR must also be accurate in the recognition of specialized vocabulary, a feature which is often neglected by most scholars, but deemed of great importance by others (for example, in Fantinuoli, 2018; and, to a minor extent, in Romero-Fresco, 2018). More specifically, advance preparation is considered one of the most important activities to ensure quality in the output of interpreters, especially in the interpretation of highly specialized domains (Kalina, 2005; Gile, 2009). According to Xu (2015), the use of coherent and accurate terminology can in fact enhance the communications, in addition to increase the perceived professionalism of interpreters. Yet, in the case of simultaneous interpretation, CAI tools offered for terminology searching may be hindered by constrains which are primarily related to time pressure and cognitive overload during the activity. Probably, the most promising implementation of ASR in simultaneous interpreting is represented by the integration of this technology directly into the workflow. Features for the automatic transcription of numbers, abbreviations, acronyms, and proper names may for example offer useful support to the interpreting effort. In fact, as highlighted by Gile (2009), these elements of speech are often a potential problem for interpreters because they imply heavy processing costs in terms of cognitive resources deployed, with severe errors and disfluencies as a consequence.

At this stage of the literature review, after examining the potential deployment of ASR technology as a CAI tool for the interpreters, it is necessary to discuss about another application of ASR technology in interpreting, i.e., the integration of the automatic transcription output generated by ASR in the interpreting process, for the direct interpretation by part of a professional. This particular configuration has been specifically examined by Fantinuoli (2017b) who, in connection with the integration in an interpreter's workstation, has provided for the identification of the following issues for ASR:

- ***"Usage of spoken language:*** *oral speech may contain different styles (formal, casual, etc.). Also in formal contexts, speakers may use spontaneous speech features, or read* *aloud prepared texts, or use a mixture of both. Speakers may make performance errors while speaking, i.e. disfluencies such as hesitations, repetitions, changes of subject in the middle of an utterance, mispronunciations, etc.*

- ***Speaker variability:*** *speakers have different voices due to their unique physical features, and personality. Other characteristics like rendering, speaking style, and speaker gender may influence the signal, together with regional and social dialects.*
- ***Ambiguity:*** *the natural language has an inherent ambiguity, i.e. it is not easy to decide which of a set of words is actually intended. Typical examples are homophones.*
- ***Continuous speech:*** *one of the main problems of ASR is the recognition of word boundaries. Besides the problem of word boundary ambiguity, speech has no natural pauses between words.*
- ***Background noise:*** *a speech is typically uttered in an environment with the presence of other sounds.*
- ***Speed of speech:*** *speeches can be uttered at different paces, from slow to very high.*
- ***Body language:*** *human speakers do not only communicate with speech, but also with non-verbal signals, such as posture, hand gestures, and facial expressions".* (Fantinuoli, 2017b: 5-6)

More specifically, the criticalities and features described above are of high relevance for the purposes of the present study, especially in relation to the first point, *"usage of spoken language"* (Ibid.). The experimental study by Fantinuoli has led to the creation of a prototype of ASR-CAI integration, the output of which has been tested in terms of output accuracy and terminological coherence. According to the study conclusions, though marketed ASR engines are still considered as not perfect and they fail under certain circumstances (non-native accents, unknown words, etc.), on the other hand, they can reach high precision values in standard conditions, even within specialized domains.

Within the context of Interpreting Studies, ASR plays an important role in the so-called Machine Interpreting (MI), also known in literature as "Automatic Speech Translation" (AST), "Automatic Interpreting" or, simply, "Speech-to-Speech Translation". By recalling the distinction between ASR and AST made above in §2.2.1, it is possible to further comment that ASR has become more and more relevant in the area of Human Language Technologies and, indeed, in correlation with Automatic Speech Translation (AST). As suggested by Satoshi Nakamura, AST has recently become *"one of the ten emerging technologies which are going to change the world"* (Nakamura, 2009: 35). To use Fantinuoli's words:

> *"Machine interpreting (MI), also known as automatic speech translation, automatic interpreting or speech-to-speech translation, is the technology that allows the translation of spoken texts from one language to another by means of a computer program" (Fantinuoli, 2018a: 5).*

Differently to what happens with computer-assisted interpreting (CAI) and remote interpreting (RI), AST (or MI) should also be considered as an important technological breakthrough contributing to the *"upcoming technological turn in interpreting"* (Fantinuoli, Ibid.: 10). De facto, as underlined by Fantinuoli (2018a: 5), AST (or MI) is a *"technology that aims at replacing human interpreters"*. When defining AST, it is evident that ASR is an important component of it. By quoting the description of AST offered by Fantinuoli:

> *"It combines at least three technologies to perform the task: automatic speech recognition (ASR), to transcribe the oral speech into written text, machine translation (MT), and speech-to-text synthesis (STT), to generate an audible version in the target language." (Ibid.).*

Within an AST system, the voice input received from the ASR system is then processed (by means of a microphone or electronic device) and elaborated. At this stage, as commented by Eugeni, *"according to the intended usage, the input can be turned into images, operations or commands or [...] in words"* (2008: 16; my translation). Under this study, the voice input produced by the speaker is turned into written text and it will be analysed like a written text. This type of process (and technology) is also denominated as *Speech-to-Text* (STT) and it incorporates two modules: a speech recognition module and a speech transcription module.

As mentioned above, AST is based on three different technological components: Automatic Speech Recognition (ASR), which is used to automatically transcribe the oral speech into a written text, the Machine Translation (MT), and the Speech-to-Text (STT) or Speech-to-Speech (STS) synthesis, which represents the last technological module in the workflow used to generate a written or audible version in the target language, respectively. Thanks to the recent developments of ASR technology, in particular the introduction of the neural networks (as seen in §2.2.2 above), the results of MI have resulted to be very promising during testing, according to Fantinuoli (2017a; 2018a), even if this technology is *"still very far from achieving the ambitious promise of a comparable quality output as human interpreters"* (Fantinuoli, 2018a: 5). In this respect, also Valentini (2002) maintained that the increasingly perfectioning of the speech recognition systems is going to inevitably lead to *"a change in the profession"* of the interpreter. Müller et al. (2016) have more recently shown the excellent, promising performance of a first prototype of MI, a real-time automatic speech translation system for university lectures implemented at the Karlsruhe Institute of Technology; other, very popular examples are the solutions offered in the market by technology giants, such as Google Translator or Microsoft (Skype) Translator. Although Automatic Speech Recognition and other CAI tools have still relatively small economic impact on the interpreting industry (as commented by Fantinuoli, 2018a), the pressure to deploy these technologies is likely to increase. As a matter of fact, as pointed out by Besnier (2012), current private and public organizations *"are obsessed with technology"* and interpreters may be asked to adopt these technologies by employers and clients. Yet the deployment of ASR technology may contribute to accelerating the process of the so-called interpreter *"depersonification"* (as commented by Fantinuoli, 2018a: 7) and, indirectly, increasing the scepticism by professionals towards new technologies.

Before discussing about the fourth group of studies on ASR and accessibility (see §2.4), the studies and theory on Machine Translation is not offered, including a presentation of the state of the art on MT. For a better understanding of accessibility studies in combination with ASR, it is fundamental to present the second most important element of the system: the Machine Translation.

2.3. Studies on Machine Translation

2.3.1. Definition of Machine Translation

A general definition of Machine Translation (also known as *Automatic Machine Translation* or *Automatic Translation* in literature) is provided by Liu and Zhang. According to these scholars' definition:

> *"Machine Translation (MT) is a sub-field of computational linguistics (CL) or natural language processing (NLP) that investigates the use of software to translate text or speech from one natural language to another. The core of MT itself is the automation of the full translation process, which is different with the related terms such as machine-aided human translation (MAHT), human-aided machine translation (HAMT) and computer-aided translation (CAT)".* (Liu and Zhang, 2014: 105)

More specifically, the authors make a distinction between pure automatic translation, which produces a completely automatic translation process, and other types of translation, which consist in an interaction between humans and machines: *i.e.*, Machine-Assisted Human Translation (MAHT), Human-Assisted Machine Translation (HAMT) or Computer-Aided Translation (CAT). The definition offered by Hutchins and Somers (1992) is different from that given by Liu and Zhang insofar as they do not define MT as a discipline:

> *"The term 'machine translation' (MT) refers to computerized systems responsible for the production of translations with or without human assistance. It excludes computer-based translation tools which support translators by providing access to on-line dictionaries, remote terminology databanks, transmission and reception of texts, etc. The boundaries between machine-aided human translation (MAHT) and human-aided machine translation (HAMT) are often uncertain and the term computer-aided translation (CAT) can cover both, but the central core of MT itself is the automation of the full translation process."*(Hutchins, 1995: 431).

As already seen in the case of ASR, these two definitions above underline that MT can be considered both as subdiscipline of computational linguistics and as a system or technology. Certainly, both definitions also highlight the fact that the systems of automatic translation deal with the conversion of natural languages, to be intended in opposition to artificial languages or programming languages. In addition, while the definition by Hutchins and Somers does not specify whether automatic translation should exclusively limit itself to written texts, the definition provided by Liu and Zhang makes it more explicit, including both the translation of written and oral texts into MT. Further definitions and details of MT are also given in sections below about the history and architectures of MT.

2.3.2. History of Machine Translation

The idea of using machines to translate natural language can be dated back to the 17th century when the concepts of universal language and "mechanical dictionary" started to circulate among philosophers and inventors (Hutchins, 2000). However, it was only during the 20th century that a true evolution of automatic translation started, with two patents by French George Artsrouni and by Russian Petr Smirnov-Trojanskij, both registered in 1933. The patent by Artsrouni provided for a sort of multilingual mechanical dictionary, while that created by Trojanskij proposed a multilingual translation device which could exploit a codification/decodification method for the grammatical functions based on the universal language Esperanto (Hutchins, 2010). According to Gaspari and Hutchins (2007), Trojanskij was a pioneer of machine translation, even if his proposal did not reach a large audience outside Russia.

It is also important to point out that the history of MT has often been strictly interconnected with the development of secret languages and codes and their decodification. For example, the decipherment of the Germans' ENIGMA code during World War II can be considered as one of first attempts to decode a secret language by means of a machine and to create a machine capable of deciphering that secret language and converting it into English. More specifically, the British team of engineers under the supervision of Alan Turing, located in Bletchley Park, was responsible for this project and it was able to break the ENIGMA code by means of statistical methods that were processed on computing machines. Those scientists and engineers laid the foundations for practical MT. From the same perspective, the euphoria about research projects on MT also continued during the Cold War period, when the threat of the Russian nuclear power contributed to large investments, mostly for the English-Russian language combination (Stein, 2018: 7).

After these first steps, the history of MT continued to be *"characterized by lows and highs, great ambitions and strong disillusionment"* (Chiari, 2007: 31; my translation). It was in July 1949 that MT started to become an object of discussion and interest for scholars in the United States and in the rest of Europe with the publication of a Memorandum (the "Translation Memorandum") by mathematician and engineer Warren Weaver. For the first time, he talked about the possibility of using a computer to produce texts from a source language to a target language. The "Translation" Memorandum (Weaver, 1949) is now considered as one of the most significant papers on the origins of machine translation, and was the result of Weaver's knowledge on cryptography, statistics, information theory, logic, and linguistic universals (as commented by Hutchins, 2010). It set forth a series of objectives and methods, stimulating research in the United States and in the rest of the world. Weaver's work had substantial influence on the highest US Government officials. On the wave of his Memorandum, in June 1952, the first Conference on MT was held at the Massachusetts Institute of Technology, Cambridge, United States, where linguists and electronic engineers joined for the first time in order to survey the linguistic and engineering problems presented by MT. At the end of the Conference, most participants had the general impression that, for certain types of texts, a mechanization of the translation process was feasible.

Another significant example of this period's research activity is the so-called IBM-Georgetown Experiment, launched on 7th of January 1954 at Georgetown University. As indicated by Riediger and Galati (2012: 7), *"with a vocabulary of 250 words only and 6 grammar rules, a selected sample of Russian phrases was translated into English"*. The experiment had such a "*large impact on public opinion as to stimulate significant financial investments on MT research in the United States and the start of similar projects in other countries, notably the former Soviet Union"* (Hutchins, 2010).

During this early stage of MT evolution, most experts in the industry agreed on the fact that the FAHQT (a "Fully Automatic High Quality Translation") principle represented almost an unachievable objective as human intervention seemed to be inevitable, given the low memory capacity/processing power of the machines of those times (*ibid.*). Later, in the 1950's, the first MT's magazine was also founded (entitled "*Mechanical Translation*"), and the first PhD thesis on automatic translation (by Anthony G. Oettinger) was published in 1955.

In general terms, during the first years of MT research, as highlighted by Hutchins 2010 (in Naldi, 2014: 60), we can maintain that three different approaches were developed:

> 1. ***"A Direct Translation model*** *– Based on a series of rules from a Source Language (SL) to a Target Language (TL), where a minimal analysis and syntactic reorganization was carried out;*
> 2. ***An Interlingual (machine) model*** *– Based on abstract language-independent representations, both from the SL and the TL. The translation is therefore carried out in two distinct phases: from SL to the interlingua and from the interlingua to TL;*
> 3. ***The Transfer-based machine translation*** *– based on three steps: analysis of SL (grammar, rules), synthesis of TL (conversion in the TL structure) and the so-called transfer modules, realizing the conversion from a language to another. In the interlingua-based MT, this intermediate representation must be independent of the languages in question, whereas in transfer-based MT, it has some dependence on the language pair involved" (Naldi, 2014: 60).*

Following these early steps and studies, with the US administration's publication of the Automatic Language Processing Advisory Committee (ALPAC) report, things radically changed as the report clearly indicated that MT could not provide for satisfactory results: it was suggested that MT was neither useful nor did it seem to provide any considerable advance or meaningful progress (Hutchins, 2010). The ALPAC report had a considerable impact on the academic world and produced a strong slowdown in MT research for over a decade, both in the United Stated and across the world.

During the 1970s, thanks to the development of early computers, research on MT undertook a new and robust evolution. In those years, new operating systems capable of running MT started to be implemented and used across the market, namely software products such as Systran, Logos and METAL. In particular, Systran was purchased by the Commission of the former European Communities in 1976, and installed into the office workstations of several intergovernmental institutions, including the North Atlantic Treaty Organization (NATO) and the former International Atomic Energy Agency or at companies such as General Motors. Although Systran, Logos and METAL were designed for non-specific vocabulary and use, during the 1970s and 1980s new systems for specific domains were also developed (Hutchins, 2010).

1989 represented an important milestone in the history of MT evolution because research started to implement new methods of MT based on corpora, that is to say large collections of texts in electronic format. Among those methods, a machine translation system based on examples (briefly, EBMT) was created: it was founded on the idea that translating often implies a process of searching for analogue examples to verify how these have been previously translated; parallel to it, another emerging system was the so-called statistical machine translation (SMT) characterized by the usage of statistical methods of analysis and synthesis and by the absence of linguistic rules. Research also continued in the field of systems which were based on rules, by both implementing the Transfer approach and the Interlingual model (mentioned above), by different groups of researchers (Hutchins: 2010).

During the 1980s and 1990s, an increasing interest for the automatic translation of spoken language also emerged and for the distribution of the first *translator's workbenches*, *i.e.*, workstations for translators which started to be launched into the market starting from 1991 (Zanettin, 2001: 27). In addition, starting from that period, the concept of MT was gradually accompanied by the concept of computer- or machine-aided translation, which distinguishes itself for the fact that it also caters for the intervention of human translators. In those years, a first distinction between *Machine Aided Human Translation (MAHT*) (where the translation is performed by a human translator, but he/she uses the computer as a tool to improve or speed up the translation process) and *Human Aided Machine Translation (HAMT*) (where the SL text is modified by a human translator either before, during, or after it is translated by the computer) was introduced, as explained in Zanettin (2001: 24).

With the arrival of the Internet in the 1990s, the technology of MT was further expanded as Web pages and E-mail contributed to increase the demand for translation and inter-communications. In particular, the main effect of this novelty was the launch onto the market of one of the first online automatic translation services, *i.e.*, Babel Fish (put online in 1997 by AltaVista search engine). This service, based on Systran system, contributed to give to the automatic machine translation visibility across the world, by making it easily accessible to Internet users. Although it was not the first online MT service, Babel Fish could distinguish itself for being open to all users, without any obligation of subscription or payment (Gaspari and Hutchins, 2007: 200).

Finally, to conclude the presentation of the history of MT, an important technological innovation was represented by the introduction of neural systems or neural networks into MT systems. In brief, Artificial Neural Networks (ANNs) or Neural Networks:

> *"…are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems "learn" (i.e. progressively improve performance on) tasks by considering examples, generally without task-specific programming". (Van Gerven, M. and S. Bohte (ed.), 2018: 5-7).*

But the architectures and components of more recent, neural MT technology will be better discussed in the next section. Here it is sufficient to highlight that the technology at the basis of neural network innovation has radically changed the MT industry, expanding the capacity and performances of MT systems. But before discussing the state of the art of NN systems, it is essential to describe the main, general architectures of an MT system.

2.3.3. The architecture of Machine Translation technology

After having presented the main steps in MT technology evolution, it is now necessary to describe the different typologies of Machine Translation technology architectures, which can be defined as: rule-based, data driven, hybrid and, more recently, a new, advanced typology, that is to say the neural or neural network-based architecture. Here below an in-depth description of these architecture typologies is provided, together with a series of critical considerations.

The **rule-based Machine Translation systems** (RBMT) represent the pioneering MT architectural solutions and it is nowadays considered as an obsolete technology. To briefly describe these systems, it is possible to underline that RBMT systems are based on dictionaries and grammar rules both for the source and the target languages to be used in the translation process. All these materials are denominated as "rules" and they are organized into different modules which interact among each other at different levels.

Developed starting from the 1990s, the **data-driven MT systems** are of more interest as they are based on statistical methods and bilingual corpora. This innovation generates the advantage of incorporating materials and texts used in bilingual corpora which have been previously created by professionals, regulatory authorizations, official institutions, etc. More specifically, these MT systems can be subdivided into example-based and statistical systems. Example-based systems are based on the concept of analogy, that is to say, matches of sequences of words are identified in the corpora, between the source and the target language. Subsequently, these matching sequences are combined together to obtain the output, that is to the target text (Hutchins, 2005). Among the data-driven MT systems, the statistical machine translation (SMT) architecture is the most dominant (Hutchins, 2005: 198) and it is based on corpora as well. Yet its functioning is not only based on matching but also on statistical probability. As a matter of fact, as explained by Chiari:

> *"[...] if we have a vast corpus of texts in the original language and in the translation version for a pair of languages (the so-called parallel corpus), we can automatically extract the most frequent matches between [...] segments of sentences or phrases. Together with matches and equivalences, we can also extract the probability by which a given segment is translated into a certain segment, rather than in a different one" (my translation: Chiari, 2011: 32).*

In this architecture configuration, the algorithms of SMT systems are "trained" on parallel corpora so that they can identify and extract translation matches that are recurrent, and thus calculate the frequency by which a given word or string corresponds to a word or string in the target text. The early SMT systems were based on a word-based approach, but the most efficient approach was the phrase-based (also known as Phrase-Based Machine Translation, or, in brief, PBMT). Within this system, sentences are subdivided and disassembled into sequences of words or phrases, which do not necessarily represent a linguistic unit (Koehn, 2009: 8) and allows for better matching of semantic units. For this reason, the PBMT is considered as the benchmark system within the SMT technology (Koehn, 2009: 8). Its functioning mechanism is summed up and described by Hutchins:

> *"Sentences of the bilingual corpus are first aligned, then individual words or word sequences (called "phrases" or "clumps" in SMT literature) of source language (SL) and target language (TL) texts are aligned, i.e. brought into correspondence. On the basis of this alignment are derived a "translation model" of SL-TL frequencies and a "language model" of TL word sequences. Translation involves the selection of the most probable TL output for each input word or phrase and the determination of the most probable sequence(s) of words in the TL (Hutchins, 2005: 198)".*

In general, it is possible to assert that, if compared to previous RBMT or EBMT systems, SMT systems offer the advantages of a better quality of the output at a semantic level, thanks to the usage of corpora containing translations carried out by professionals, with a higher level of accuracy (Hutchins, 1995). On other hand, for a high quality output, these systems need to use large, high-quality corpora, as well as the necessity of meeting high-capacity hardware requirements in terms of computational processing capacity for the management of large translation models. Yet, in the case of large organizations or international institutions, the technology and memory requirements of these systems are often met.

The combination of rule-based and data-driven systems led to the creation of the so-called **hybrid systems**, in the attempt of overcoming the limits of the two previous typologies. In fact, RBMT systems require for *"explicit rules of lexical, morphosyntactic and semantic nature allowing for the transfer from the input to the output"* (my translation: Gaspari, 2011: 24-25), as well as a complex, time-consuming processing, both in linguistic and in computational terms. On the other hand, SMT systems permit to design a complete, functioning system in rapid times, but they *"are characterized by a sectorial specificity which derives from the textual typology and the sector of the parallel corpora used for training these systems"* (my translation: Gaspari, 2011: 26). The solution to these criticalities is represented by hybrid systems which offer and try to exploit the strengths of the rule-based systems and of the corpora-based systems, while combining statistical methods with linguistic rules.

As seen in §2.3.1 about the history of MT technology, a recent, most advanced innovation is represented by the so-called **neural networks-based MT systems** or

Neural Machine Translation (NMT) systems, which have revolutionized the industry of MT. To better clarify this concept, NMT is a technology based on an artificial network of neurons. In the last years, this technology made significant progress thanks to Artificial Intelligence (AI) and it is now widely used as a starting point for professional translation services (as reported in *SDL Research Survey 2016* (SDL, 2016). NMT allows translating in real time information and documents with accuracy and reliability levels which may be possibly compared to those of professional translators (although a large debate is still ongoing about this consideration among scholars). Examples of its market application are automatic translation software solutions such as Google Translate, Microsoft Translator or DeepL (to name just a few of them). By using the definition offered by Starnoni, NMT can be described as follows:

> *"A neural network is an artificial representation of the knowledge composed of thousands of units, or nodes, whose functioning systems gets inspiration from that of human neurons. Each of these nodes is associated to a given concept and it is in a precise position, which can be identified by means of vectors".* (Starnoni, 2019)

These networks are therefore mathematical/IT models which are designed to emulate the behaviour of neurons in human brain. De facto, like a biological neural network, the NMT system receives external data and stimuli which are processed by a huge quantity of interconnected neurons, artificial neural networks are capable of modifying their nodes according to the external and internal data. More specifically, in neural networks, the knowledge required to carry out a specific task is distributed across all the nodes in the network, as shown in Figure 2.3 below.

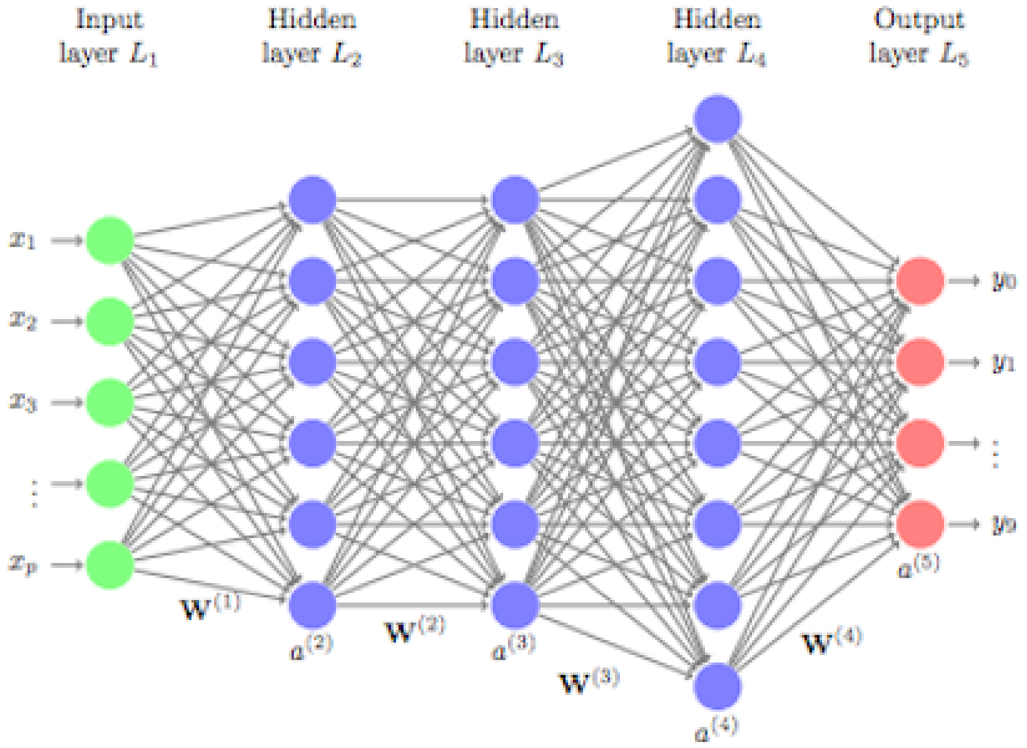


Figure 2.3 – Architecture of Deep Neural Network-based MT[16]

Within the NMT system, neurons are distributed across lines which are denominated as "layers": the system thus incorporates layers of input (collecting the incoming data), one or more intermediate layers (also known as "hidden layers") and the layer of the output, which provides for the results. When hidden layers are two or more than two, the network is defined as "deep" (hence, "deep neural network") or "deep learning" networks. To use the words of Deng and Yu, *"deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input"* (Deng and Yu, 2014: 199-200). For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces. In line with these considerations, Wu et al. describe deep learning NMT systems as an *"end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems"* (Wu et al., 2016: 1). From Castilho et al. (2019: 1) it also possible to learn that *"over the past five years the Machine Translation (MT) community has become aware of the potential of Neural Machine Translation (NMT)"*. In particular, Kenny (2018) highlighted that this

[16] Source: https://afit-r.github.io/feedforward_DNN

technology contributed to *"increases in output quality that had appeared to plateau when using statistical MT (SMT)"*. As also commented in Castilho et al. (2019: 1), *"early studies on NMT quality demonstrated that, in general, this MT paradigm yields higher automatic evaluation metric scores than its predecessor, SMT"*. NMT has also been shown to provide greater fluency when compared with SMT (as underlined by Bentivogli et al. 2016; Toral and Sánchez-Cartagena 2017). This also explains the adoption of this system by several multinational translation agencies and international institutions.

When examining the nature and composition of NMT systems, it is possible to find out different typologies of neural networks. Examples are the recurrent neural networks and the so-called feed-forward neural networks. The latter is a typology of neural network in which information moves towards one direction only: from input nodes to intermediate nodes (if existing) and then to output nodes. In recurrent neural networks (RNNs), in addition to the ascending connections of feed-forward networks, also descending (denominated as "recurrent") connections are established, connecting output units to intermediate and input units (Cho et al., 2014). Recurrent networks can adjust their outlooks according to the previously processed data, a process similar to that of "learning" and improving thanks to the data entered.

The typical architecture of a neural MT system is composed of two RNNs denominated as encoder and decoder. During the training phase of the neural system, bilingual corpora are used: the encoder transforms the input text into a vector which is then turned into the output text by the decoder. This operation of transformation and "correction" is repeated until the system reaches the best possible results. The mechanism is well represented in Figure 2.4 in the next page. The figure below and the mechanism of a neural network can certainly be better understood by using the description offered by Starnoni:

> *"During the first phase, the encoder creates a representation of each single phrase in its context, by breaking up each sentence of the initial text. Each of these representations is merged with that of the following word, creating a new representation: this process is applied on a repeated basis generating outputs which are re-used from time to time. The system learns to remember only the outputs which are useful and relevant, while forgetting the other ones. During the second phase, the decoder assigns to each representation a series of words that, with a certain degree of probability, constitute the correct continuation of what was previously written, on the basis of both the position of the word*

> *in the destination sentence and of its relations within the destination linguistic code" (my translation: Starnoni, 2019)*

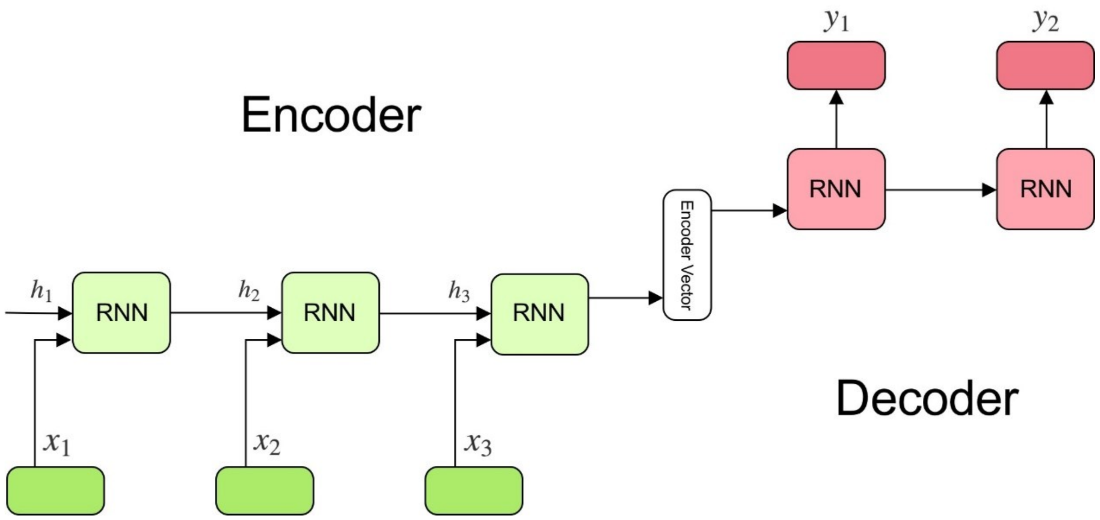


Figure 2.4 – Recurrent Neural Network architecture and functioning.

Given that these systems are capable of positioning words that are semantically similar at a short distance from each other within the vectorial space, NMT systems can capture the semantic content of sentences in a more efficacious way. Additionally, NMT systems are capable of taking into considerations the textual references within the sentence by identifying the distant references and improving fluency (Starnoni, 2019; Cho et al., 2014). This feature has been further enhanced with the introduction of the so-called "attention mechanism" into the RNN, which is able to "suggest" to the decoder which part of the source text must be taken into consideration during the generation of each target word. The mechanism can be described by quoting Bahdanau et al.:

> *"The decoder decides parts of the source sentence to pay attention to. By letting the decoder have an attention mechanism, we relieve the encoder from the burden of having to encode all information in the source sentence into a fixed-length vector. With this new approach the information can be spread throughout the sequence of annotations, which can be selectively retrieved by the decoder accordingly (Bahdanau et al., 2015: 4)".*

Thanks to the introduction of the features described above (the recurrent mechanism and the attention mechanism), RNN systems have significantly improved their performances. This has paved the way to the creation of a new type of architecture, which is denominated as "Transformer". In particular, Vaswani et al. (2017) have described this recent, innovative architecture developed by a team of Google researchers. Most of today's state-of-the-art RNN systems (for example, Google Translate) are actually based on that architecture. The structure of the Transformer architecture incorporates a encoder-decoder system. In turn, each encoder incorporates a self-attention mechanism and feed-forward, while the decoder module incorporates (in addition to the self-attention mechanism and the feed-forward features) the encoder-decoder attention mechanism, which carries out the same function of the previously described feature. Figure 2.5 below graphically present the Transformer architecture:

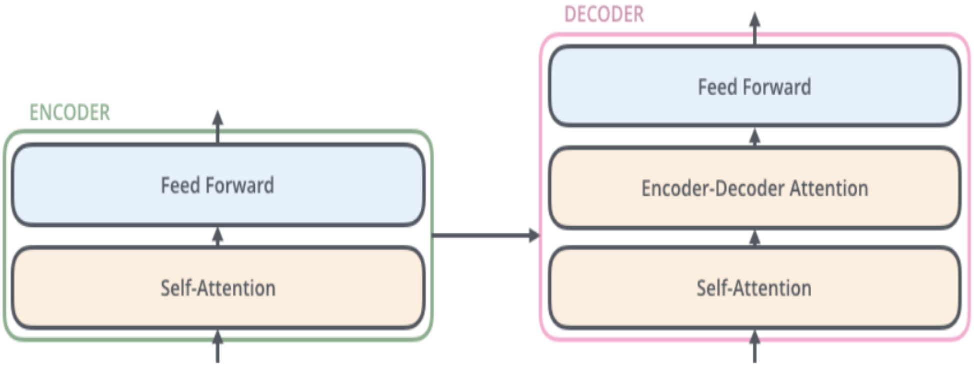


Figure 2.5 - The "Transformer" architecture in current, most advanced NMT technology

One of the key, innovative aspects of the Transformer model is indeed the self-attention mechanism which permits the encoder to "look at" the other words which compose the input sentence while it generates the vector relating to a word of the same sentence. In other words, the self-attention mechanism keeps into consideration the relation between an input word and all other words in the source text more efficaciously; if, for example, interdependence relations across the words that are part of the input exist, the mechanism is capable of "capturing" the strong bond across these words and to preserve it in the encoding phase.

In recent years, several scholars have underlined the improvement in terms of performance and quality in the usage of RNNs, in comparison to SMT systems or other previous MT systems. For example, Bahdanau et al. (2015) and Sutskever et al. (2014) have examined the quality of NMT systems in comparison with PBMT systems by using automatic evaluation metrics such as the BLEU value, with equivalent or higher performances on neural networks. These early works have then been followed by other research groups and single studies, like for example Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Van Brussel et al., 2018; Klubička et al., 2017; Castilho et al., 2018a, to mention just a few of them.

Another important aspect of discussion within the scientific literature on MT technology is represented by quality evaluation. In particular, the management of quality evaluation can be carried either by humans or automatically. In the first case, a certain degree of subjectivity is reported and this has taken to a wider adoption of automatic evaluation systems of MT output across most of the reviewed scholars. Under the automatic approach, the comparison between a "gold standard" and the MT results is often at the basis of every evaluation process. Generally, the so-called gold standard is represented by the corresponding translation of human translators or professionals, who are considered as the "benchmark" for the evaluation (Castilho et al., 2018b). This approach is epitomized by the BLEU metrics, according to which *"the closer a machine translation is to a professional human translation, the better it is"* (Papineni et al., 2002: 311). More specifically, the BLEU metric is an automatic evaluation method developed by IBM in 2002 and today it represents the benchmark for most of MT quality evaluation studies (Castilho et al., 2018b: 26). BLEU is mostly based on the concept of precision, according to which this metric calculates the number of n-grams of different lengths shared between the MT-generated output and the reference translation. BLEU expresses the proportion between the number of n-grams of different lengths (typically from 1 to 4) that appear in the MT output and in the reference translation; this value is then divided by the total number of n-grams of that specific length in the output. In addition to the principle of "precision", BLEU is also based on the principle of "recall", which is the proportion between the number of correct words contained in the output and the number of total words in the reference translation (Koehn, 2009: 223).

Among the most popular methods of automatic quality evaluation there are probably the WER and TER rates, in addition to the BLEU metric. The Word Error Rate (WER) was firstly originated and derived from the Automatic Speech Recognition industry, as better described in Section 2.4. on Accessibility Studies. As highlighted in Castilho et al., 2018b, the WER rate in MT quality evaluation is obtained by calculating the number of additions, omissions and substitutions necessary for the perfect matching between the MT output and the reference translation output. The lower the WER rate, the better the quality of the MT output. Similarly, the TER (Translation Error Rate) metric calculates the number of editions which are required to carry out for obtaining a perfect matching between the MT and the reference translation outputs; the value so obtained is then normalized with respect to the reference sentence length (Snover et al., 2006). Differently from what happens with the WER rate, the TER rate also considers the shifts of words or sequences of words. The TER rate is between 0 and 1: a value closer to 1 implies a higher number of editions and thus a lower quality output. As highlighted by Papineni et al., *"few translation will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1"* (Papineni et al., 2002: 315). These considerations will be particularly relevant in choosing the evaluation metrics to be used in the analysis of the NMT system in the present study's pipeline.

2.4. Accessibility Studies

The right to accessibility and media accessibility are pivotal concepts for all accessibility studies and projects, as commented in Greco (2016: 1) and in Romero-Fresco (2018: 188). While previous projects and studies (see §2.2.3.2 above) were mostly aimed at meeting the needs of institutional organizations (*e.g.*, TC-STAR, EU-BRIDGE) or services (*e.g.*, DARPA-GALE), accessibility studies and, more specifically, media accessibility studies focus on the use of assistive technologies for the purposes of breaking down the communication barriers for physically-impaired people or individuals with physical disabilities, for example the non-hearing people. The concept of accessibility as a universalistic right stemmed from the regulatory framework set up by the United Nations and by the European Union institutions in the course of the Twentieth and Twenty-first centuries. In particular, the very concept of accessibility derives from the *Universal Declaration of Human Rights* (UDHR) of the United Nations (Paris, 1948) and from the General Comments (page 5) of document E/C.12/1999/5 published in 1999, in which the UN Committee on Economic, Social and Cultural Rights highlighted the role of each single state or national government: *"the State must pro-actively engage in activities intended to strengthen people's* *access to and utilization of"*. More recently, the right to accessibility was certainly spurred by the approval of the UN *Convention on the Rights of Persons with Disabilities* (CRPD) of 2006. In particular, the *General Comment on Article 9* of the CRPD – released by the UN Committee on the Rights of Persons with Disabilities in 2014 – represents a milestone in the international disability movement to establish a new interpretation of disability and of persons with disabilities within society, where accessibility is considered as a fundamental right (as already mentioned in the Introduction to this thesis). This aspect is of significant relevance for the accessibility and media accessibility studies because institutions are now required to provide for the "resources" necessary to guarantee accessibility to communication. Quoting Greco (2016: 2), *"assessing whether accessibility is a human right per se (or if not, then defining what exactly it is) is of the utmost importance for the field of human rights, as well as the struggle for inclusion of persons with disabilities*".

Narrowing the scope of this general review, it is worth mentioning here those regulatory and normative provisions, within the EU and in the industry of audiovisual production, which contributed to the definition of the "objects" and resources of accessibility mentioned above and, more specifically, the definition of the concept of Media Accessibility (MA), which is a fundamental pillar of the present study. In fact, the object of this study (*i.e.*, the subtitles generated by ASR) is a form of digital media. The *EU Audiovisual Media Services Directive (2016)* has identified the provision of MA as a *"necessary requirement not only for persons with sensory impairments, but also for older people to participate and be integrated in the social and cultural life of the EU"* (Romero-Fresco, 2018: 188). The latest *International Standards on Subtitling, ISO/IEC DIS 20071-23* (International Organization for Standardization, 2018), cites as its main target users *"persons with hearing loss, persons who are deaf or hard of hearing, persons with learning difficulties or cognitive disabilities"* among others, as reported in Romero-Fresco (2018: 188).

Many of the reviewed works conducted on the theory and practice of MA have generally focused on access to audiovisual content. In particular, the target groups of this access have been the deaf and blind communities, as claimed by Romero-Fresco (2018: 190). On the other hand, in recent years, the literature on this field has concentrated its attention on two emerging areas: *i.e.*, interlingual respeaking and accessible filmmaking (AFM). These subfields of research have set forth:

> *"The need to open the scope of MA to other groups, including the elderly, children, people with learning disabilities and people without disabilities who may need linguistic access to audiovisual content in a foreign language"* (Romero-Fresco, 2018: 190).

As far as the studies on accessible filmmaking are concerned, here it is necessary to specify that, given the scope of the present study, those studies are not of particular relevance as they focus on the *"consideration of accessibility during the production of audiovisual media in order to provide access to content for people that cannot, or cannot properly, access it in its original form"* (*ibid*: 192). The object of these studies is therefore different from the subtitling output of ASR examined here.

As introduced in §2.2.3 above, the studies on ASR also include a group of studies focusing on the use of ASR technology for accessibility purposes and for the subtitling industry; for convenience, it has been decided to present those works under this section. This wide range of works is presented here in order to depict the current state of the art in terms of subtitling standardization and quality evaluation. The goal is to highlight the impact of ASR technology on accessibility improvement and on generating subtitling for Media Accessibility (MA) and Audiovisual Translation (AVT).

To describe what monolingual (or intralingual) and intralingual subtitling is about, it is possible to use the definition offered by Caimi, who describes it as a *"form of screen translation which involves the transfer from oral language into written language"* (Caimi, 2006: 86). Subtitling can in fact serve both as an accessibility aid for non-hearing people, but also as supplementary aid for different purposes (for example, for second-language learners). Under the latter case, subtitling is defined as **didactic aid**, by quoting Caimi. Probably, the distinguishing feature of subtitles as an **accessibility aid** is represented by its supplementary and complementary nature. More precisely, as underlined by Caimi, *"it is the intentional combination of the phonological expression of the foreign language with its written form that acts as a complementary aid to language comprehension"* (Caimi, 2006: 87). In a simpler way, Jakobson defined intralingual translation or subtitling as the *"interpretation of verbal signs by means of other signs of the same language"* (Jakobson, 1959: 233). The primary aim of intralingual subtitling is to cater for the needs of the deaf and the hard-of-hearing.

For the purposes of the present study, the most important topic of discussion within Accessibility Studies is probably the definition of quality and, more recently, the discussion about the impact of ASR technology on subtitling production. In this respect, over the past years, an increasing number of publications (for example, Neves, 2018; Remael, et al., 2012) on audiovisual translation (AVT) and, especially, on media accessibility (MA) have pointed out that the focus is significantly shifting from quantity to quality. This approach has also been confirmed by other key players and stakeholders in the industry of subtitling, i.e., accessibility service providers, user associations and governmental regulators. Yet it should be remarked that less consensus is achieved among the different scholars and players concerning the modality in which quality should be evaluated. In most of the literature reviewed for the purposes of the present study, there is a significant difficulty in establishing and agreeing on what quality really means. Quoting Pedersen, *"quality is about as elusive an idea as happiness, or indeed, translation"* (2017: 210). In the translation industry, especially from an academic perspective, it is often a question of *"equivalence and language use"* (Ibid). As Pedersen additionally observes (Ibid.), *"many people have to judge translation quality on a daily basis: revisers, editors, evaluators, teachers, not to mention the subtitlers themselves, and of course: the viewers"*. And, in order to evaluate quality, assessment methods are required. A second difficulty thus emerges in addition to the definition of what quality means: that is to say, finding models that can be accepted by all stakeholders and models which can offer comparable results.

As already partially mentioned above under the current section, an emerging approach to Media Accessibility and subtitling for the subtitling industry has been calling for a wide, universalistic view of media accessibility which is focused not only on individuals with sensory disabilities, but also on anyone who cannot or cannot completely access audiovisual content in its original form (as discussed in Greco, 2018; Pablo Romero-Fresco, 2018a). This approach is mostly based on the EU Audiovisual Media Services Directive (2016), which is targeted at both persons with sensory impairments and older people. Another important framework of regulation for the subtitling industry is represented by the latest international standard on subtitling, i.e. the ISO/IEC DIS 20071-23 (Standardization, 2018), which indicates as its main target users not only persons with hearing loss, but also the individuals with learning or cognitive difficulties, persons who cannot hear the audio content due to environmental conditions (for example, noisy surroundings or circumstances where the sound is not available or not appropriate), as well as persons watching a movie in a non-native language (the didactic aid indicated by Caimi and mentioned above).

Within the field of Media Accessibility and subtitling, a key concept for the present study is that of translation quality assessment for interlingual accessibility purposes (from a source language to a target language), which has traditionally been an issue of debate (House, 2009). As explained by Doherty (2017: 131), translation quality assessment aims to *"ensure a specified level of quality is reached, maintained, and delivered to the client, buyer, user, reader, etc., of translated texts"*. Apart from being important for translator training and professional certification, this process is of the utmost importance for the evaluation of the final quality of Media Accessibility and audiovisual translation as it also allows for an analysis of the performance of translation technologies, such as machine translation or Automatic Speech Recognition. In general, within the translation studies, the translation quality assessment process has been approached *"from a theoretical and case study perspective"* (ibid:132), focusing on analysis and comparison between source text and target text (SL-TT equivalence), as well as with challenges such as subjectivity or user perception (for example, in Bassnett-McGuire, 1991; Bowker, 2000; Koponen, 2012; Snell-Hornby, 1992), lack of systematic approaches (Bassnett-McGuire, 1991) and inconsistency in terminology (Brunette, 2000). The main issue of translation quality assessment is, according to Doherty, *"the lack of explicit operationalization of concepts"* and the *"non-adherence to established standards upheld in test theory, namely those related to validity, reliability, and the selection of evaluators"* (2017: 132). Generally, the concept of reliability coincides with the degree of consistency of the test results across different evaluators (Bachman and Palmer, 1996; Clifford, 2001). And this notion is important for quality assessment as it leads to the notion of inter-annotator agreement and to the importance of the selection and training of evaluators.

In practical terms, the studies on Media Accessibility with translation quality assessment as the main focus offer useful hints and considerations for the purposes of this study, in particular for the analysis and evaluation of ASR and NMT output. More specifically, according to the literature reviewed, the analysis and evaluation of outputs can involve both a human and a (semi-) automatic assessment. From the debate among scholars, it is evident that human evaluation can probably offer the benefit of rationale judgement, but, on the other hand, it has the disadvantages derived from subjectivity (as highlighted in Koponen, 2012) and time (as commented in Doherty, 2017). By contrast, automatic (or semi-automatic) evaluation with metrics such as BLEU (Papineni et al., 2002), and TER (Snover et al., 2006) has proved to be certainly less time consuming, but it does not allow for sophisticated judgements about specific elements of speech language such as idiomaticity, naturalness, etc. Starting from a dedicated research on this matter, Doherty et al. (2013) pointed out that translation quality assessment in Media Accessibility and in the subtitling industry, in general, tends to adopt a combination of human evaluation and semi-automated methods that may or may not achieve a previously set quality threshold. In other words, the main methodology is probably that of comparing the results with a gold standard system.

For the methodology of analysis and evaluation of results, the background literature also offers two important requisites to be met. In fact, the methodology adopted should follow a model which has to be "rigorous" (research-informed, valid, reliable, user-focused) and "transferable" (straightforward, flexible and valid for training), as commented in Romero-Fresco (2020). As seen above in this section, the problem of subjectivity is often at the heart of the debate on quality assessment within translation studies and, at the same time, within Media Accessibility and the subtitling industry. This issue can be coped if the model adopted in the evaluation is rigorous or, more specifically, if it is research-informed, valid, reliable and user focused. In particular, if a model is research-informed (i.e., based on previous research), *"it may help to dispel the fears of subjectivity that are often attached to what are regarded as prescriptive models based on the individual experience of the researcher"* (Romero-Fresco, 2020). For example, considering one of the most widespread models of quality assessment in subtitling for Media Accessibility, the NER model, it is possible to assert that its formula is derived and mostly based on the basic principles of WER (word error rate) model, as applied by the US National Institute of Standards and Technology and on its adaptation by the Centre de Recherche Informatique de Montréal (CRIM) (as explained in Pablo Romero-Fresco, 2016). Also with respect to the classification of errors in terms of severity (minor, standard and serious), it is possible to underline that the model is based on the research project set up in 2010 by the Carl and Ruth

Shapiro Family National Center for Accessible Media (Apone et al., 2010) and especially on the findings of the EU-funded DTV4ALL project (Romero-Fresco, 2015).

As far as the requisite of validity is concerned, according to Romero-Fresco, the *"model must measure the dimensions that determine quality in the specific MA modality at hand"* (Romero-Fresco, 2020). So, for example, for the NER model, the parameters and dimensions which are measured (i.e., accuracy, speed and delay), are agreed on the basis of official consultations by governmental regulators in the UK and Australia with broadcasters, subtitling companies, researchers and user associations (Ofcom, 2015). Yet in the assessment of accuracy a certain degree of human intervention is required to verify, for example, if a loss of information should be accounted for in the evaluation of the final results. Hence the need to adopt a "remedy" that can mitigate the degree of subjectivity introduced by such human intervention, that is to say the need for meeting the requisite of reliability.

Across the studies on accessibility (and respeaking), in the literature reviewed for the purposes of the present thesis, a key element for the reliability of a model of quality assessment is certainly the calculation of the inter-rater or inter-annotator agreement rate between different evaluators, who, prior to this, must be selected and trained. An example of this is offered by the Live Respeaking International Certification Standard (LiRICS) initiative, the first official certification of respeakers, in which the assessment was carried out by a team of NER-certified external evaluators belonging to the research group GALMA (Galician Observatory for Media Accessibility).

As already mentioned above, a rigorous model for the quality assessment in MA and in the subtitling industry is also expected to be user-focused. The user-centred approach has traditionally been an important issue of translation quality assessment (as underlined in Ray et al., 2013). This requirement is set in accordance with the second of the three shifts produced by the accessibility revolution according to Greco: *"the change from a maker-centred to a user-centred approach"* (Greco, 2018). The role played by the raters/annotators in the case of the NER evaluation model used in the scientific literature (see Chapter 3) is represented by the different degrees of error severity (and thus the final score) assigned to each error. In other words, the model becomes user-focused because it measures the impact that an error may have on the raters and possibly on final users (though their evaluation may not match). The score is then based exclusively on the experience of the rater/user. After this consideration, the raters or annotators under the present study may be considered, to a minor extent, as the final users of the ASR output to be examined.

The rigorous requirement (research-informed, valid, reliable and user-focused) can certainly contribute to guaranteeing that a model of quality assessment is solid and as objective as possible. However, it may not be sufficient to obtain an impact on society, that is to say to have a certain utility. In fact, according to Romero-Fresco (2020), for a model to become useful in the subtitling industry and for the purposes of MA, it also needs to be transferable: that is to say, *"straightforward, flexible and valid for training"*. The necessity of combining these needs with those of rigour may certainly imply difficult decisions for the researcher or the subtitling project producer/editor, who may need to simplify elements of the model to make it more accessible for the evaluators without compromising its rigour. As a matter of fact, the necessity of being coherent with previous research and of being a valid and reliable model may imply complex methods which can prove too complicated or time-consuming for the subtitling industry or an organization using the model to apply it regularly. Hence the need of determining and defining a straightforward model. For example, reducing the number of error classifications or the levels of severity can help in replicating and transferring the system across different organizations or situations. By contrast, Eugeni (2008) prefers to promote a more complex model which can determine and identify more detailed causes and types of errors in live subtitling. But a wide array of error classifications can actually hinder the understanding of the evaluation process. As already specified above, a simpler model can actually offer the advantage of being relatively easy to understand, and this aspect is of utmost importance in the case of large-scale projects, where it is necessary to train a high number of evaluators. Certainly, simple models can generate a large amount of efforts as well. For example, the WER and NER models are both based on the comparison between the original audio and the subtitles and both need a transcription of the source speech to be carried out and analysed (gold standard): this operation determines significant efforts in terms of time and costs for obtaining an efficient evaluation (Romero-Fresco, 2020). Finally, the very concept of "straightforward model" also implies the possibility of having simple results, which can be easily readable by part of other users or evaluators; to quote Romero-Fresco (2020), the results which a model produces *"should be measurable and recognizable"*.

For a model to be transferable, it should also be flexible: the model adopted should allow for the possibility of adapting its classification to the local context (Romero-Fresco, 2020). It is therefore essential for any MA or subtitling project or study, to weigh up the benefits of adopting a standard, consolidated model capable of producing comparable results. When considering transferability, the training of evaluators or subtitlers is also equally important. In fact, in order to be transferable, the model has to be *"valid for training"* (Romero-Fresco, 2020). This requisite provides for the methodology implemented the possibility of improving the expertise or skills of evaluators or subtitlers, in the course of the time, with practice ("training"). If a model is valid for training, the possibility of improving the quality and performances of evaluators contributes to reach higher quality also in MA final product: i.e., the subtitles. Thanks to a daily-usage of WER and NER models for the evaluation of subtitles, organizations or companies deploying these methodologies may probably offer training to their evaluators and also a certain degree of consistency. Ultimately, this continuous operation can contribute to the comparability of the results.

Within the studies on respeaking, particular relevance should be given to those works that have analysed the interaction and combination of respeaking with Automatic Speech Recognition, in particular the studies where the efficiency of respeaking is compared to that of manual transcription and of automatic speech recognition (Sperber et al. 2013, Bettinson 2013). Al-Aynati and Chorneyko (2003) found that ASR-based transcriptions can become an efficient tool to transcribe medical reports, but, paradoxically, it was proved that this process required more time than manual transcription because of the extra time needed to correct the errors caused by speech-recognition software in a highly-terminological context like the medical one, thus highlighting the relevance of terminology. Within the European context, two EU-funded projects denominated "Translectures" and "SAVAS", respectively, explored the usage of ASR to improve the efficiency of transcription and subtitling for the purposes of accessibility (but not limitedly to that scope). More specifically, Translectures was focused on the development of a series of tools for the automatic transcription and the translation of online educational videos (the so-called "didactic aid" defined by Caimi, 2006). The results of the various projects conducted under this initiative concluded that the automatic generation of subtitles through ASR, plus a manual review process to eliminate errors, proved to be considerably faster than the traditional manual production of subtitles (Valor Miró et al. 2015). The SAVAS project was aimed at enhancing and realizing an ASR-based solution into seven languages (Basque, Spanish, Italian, French, German, Portuguese and English) for the production of fully automatic (and respeaking-based) subtitles and transcriptions. As reported in Álvarez et al. 2015, also the SAVAS experimentation offered promising results both in terms of accuracy and efficiency, when compared to manual transcriptions. In this case, the accuracy evaluation was mostly based on accessibility considerations and involved the respeaking process.

In general, it is possible to maintain that the studies on interlingual respeaking raise a series of challenges and share many aspects and issues with the present study, where speech recognition is involved in the generation of subtitling. In fact, as specified in Romero-Fresco (2018: 191), *"respeaking is a modality of MA concerned with the production of (live) subtitles through speech recognition"*. Yet it should be commented that this kind of studies represents a very limited portion of the entire study production on MA because they are quite recent. In fact, *"only 4% of the academic publications on accessibility and 0.8% of published outputs on audiovisual translation (AVT), respectively, deal with live subtitling and respeaking"* (Ibid: 191). Among the most relevant projects on interlingual respeaking and the usage of speech recognition is the Interlingual Live Subtitling for Access project (ILSA, 2017-2020). Promoted by the European Union and conducted by four European universities (University of Vigo, University of Vienna, University of Warsaw, University of Antwerp) between 2017 and 2020, this project has significantly contributed to identify the skills required for the professional profile of a respeaker and of a live subtitler. The project has in particular emerged from the necessity of responding to the needs of a wider audience of physically-impaired users. In fact, despite an increase of MA subtitling (especially in the filmmaking industry), the *"narrow view of MA as including mainly people with hearing loss has proved to cater for hard-of-hearing people more than for deaf people"* (Romero-Fresco, 2018: 192). In many live situations, as signers cannot also act as interpreters in a foreign language, deaf people are often forced to use subtitles that have been designed for hard-of-hearing viewers. These subtitles are often *"fast, near-verbatim subtitles that have shown to pose comprehension problems for many signers who read them in what is effectively their second language"* (Romero-Fresco, 2016). The deaf minority is thus to some extent left behind for the benefit of a majority of hard-of-hearing viewers. It is therefore of utmost importance to produce accurate subtitles and *"ensure that wider access does not involve lower quality"*, as highlighted by Romero-Fresco (2018: 192). The most interesting outcomes for the purposes of the present study is indeed the methodology of research adopted in the various research subprojects conducted within the ILSA framework. For example, the implementation of a quantitative approach in the evaluation of accuracy (see the NER, WER in Chapter 3) of ASR technology's output is of absolute relevance, together with the different considerations to be made regarding the communication scenario when non-hearing and/or deaf users are involved. But like most of MA studies, also the ILSA project did not examine the impact of Terminology on the production of subtitles and their terminological accuracy.

Another interesting project focusing on the combination of respeaking and ASR is the Shaping Multilingual Access though Respeaking Technology (SMART) project. This ongoing multidisciplinary international project focuses on interlingual respeaking (IRSP) for real-time speech-to-text and tries to address key questions around IRSP feasibility, quality and competences. The pilot project is based on experiments involving 25 postgraduate students who performed two IRSP tasks (English-Italian) after a crash course. In addition to statistical metrics, this project also involves the application of quantitative measurement. In fact, the analysis examines subtitle accuracy rates by comparing the results with participants' subjective ratings and retrospective self-analysis. In the preliminary results, as explained by Davitti and Sandrelli (2020), when commenting on the utility of ASR technology, participants have indicated multitasking, time-lag, and monitoring of the speech recognition software output as the main difficulties. The final results of SMART have not been published yet (probably available in 2022).

At this stage, as the purpose of this study's analysis is that of evaluating intralingual and interlingual (ASR and NMT) output in live conferences held at international institutions, after having examined the most relevant studies and projects on accessibility and media accessibility, it is now necessary to briefly present the main reference studies on Institutional Translation, including the role and function of translation and interpreting services within the institutional scenario.

2.5. Studies on Institutional Translation

For a definition of Institutional Translation, it should firstly be highlighted that there is no uniform understanding of its defining features or field of application, as commented by Koskinen (2014: 479). To formulate a general definition of it, institutional translation can be defined as *"any translation carried out in the name, on behalf of, and for the benefit of institutions"* (Gouadec, 2007: 36). As commented in Schäffner et al. (2014: 493), *"in the widest sense, any translation that occurs in an institutional setting can be called institutional translation, and consequently the institution that manages translation is a translating institution"*. In literature, the concept of institutional translation generally refers to translating in or for a specific organization (Kang, 2008: 141).). Koskinen provides a more detailed definition of this concept:

> *"[We] are dealing with institutional translation in those cases when an official body (government agency, multinational organization or a private company, etc.; also an individual person acting in an official status) uses translation as a means of 'speaking' to a particular audience. Thus, in institutional translation, the voice that is to be heard is that of the translating institution. As a result, in a constructivist sense, the institution itself gets translated". (Koskinen, 2008: 22)*

Accordingly, under this perspective, it is possible to assert that all institutions may produce translations, but not all of them necessarily produce institutional translations.

The first attempt of interconnecting the role of institutions with that of translation was probably made by Brian Mossop who argued that translating institutions are a *"missing factor in translation theory"* (Mossop, 1988: 65). However, it was only 20 years later, in 2009, that the concept of *"institutional translation"* (Kang, 2009) was included as an entry in the Routledge Encyclopaedia of Translation Studies (2012, 2nd ed., ed. by M. Baker and G. Saldanha). Thanks to the studies by Koskinen (2000 and 2008), the relationship between translation and institutions (e.g., the European Commission) gained the interest of scholars and contributed to the visibility of this subfield of research. Additionally, as commented by Kang (2014: 471), *"one shared theme that is evident across the collection of papers is that translation in institutions demands its own institutional frame of reference"*. To use the words by Koskinen (2008: 17), within the context of translation studies, an institution can be broadly understood as *"a form of uniform action governed by role expectations, norms, values and belief systems"*. In particular, this scholar examines the complex conceptual problems related to institutional translation and explores *why* rather than *what* institutional translation is. In fact, as underlined by Kang (2014: 471), she argues that *"translation is employed in multilingual institutions for its governing function and that the role of translation in governance is historically determined"*. Finally, according to Koskinen:

> *"The combined process of governmentalisation, multilingualization and globalisation enhances the need for institutional translation, and that rather than viewing institutions, government, and institutional translation as stable and fixed entities, it is important to examine the processes and historical trajectories through which they emerge" (Kang, 2014: 471).*

Consequently, the view of Koskinen is radically different from that of Gouadec (2007). In fact, for Koskinen, institutional translation is one of the results of the process of institutionalisation rather than being merely a service located within and serving particular institutions at some point in time (Gouadec's view).

An important element in institutional translation definition and analysis is certainly discourse. In fact, to use the words by Kress (1995), *"institutions are social constructions that are constituted through discourse"* (in Kang, 2014: 474). Provided that this study intends to examine speech transcriptions, discourse analysis is therefore of primary relevance. From this perspective, as Fairclough points out, an institution is:

> *"an apparatus of verbal interaction, or an "order of discourse", characterized by its own set of speech events, its own differentiated settings and scenes, its cast of participants, and its own norms for their combination" (Fairclough (2013: 40).*

Hence, translation in institutions, as commented in Kang (2014: 474), *"is a discursive phenomenon that involves the shifting of discourses across institutional, as well as linguistic, boundaries"*.

Another important aspect worth discussing with regard to institutional translation is the relationship between multilingualism and translation. Multilingual institutions deal inevitably with translation issues in order to maintain and protect their multilingual nature, while institutional translation depends deeply on how multilingual the institution is. As highlighted by Meylaerts, *"at the heart of multilingualism, we find translation"* (2010: 227). Institutional translation attracted academic attention from many different perspectives, both in Translation and Interpreting studies. They range from the analysis of (un)official interpreting practices in public service contexts (Angelelli, 2004; Antonini et al., 2017), to the analysis of translation policies and practices in multilingual regions and organizations (Branchadell and West, 2005; Meylaerts, 2011; González Núñez, 2014; Schäffner et al. 2014), to the use of AI technology for translation and interpreting services (namely, Machine Translation and Automatic Speech Recognition).

If one thinks of multilingual institutions like the European Union, the Swiss Confederation or the Canadian Government, it is possible to discover that translation is one of their driving forces. In such institutions, professional translators and interpreters are responsible for the preservation of institutional multilingualism. However, professionalism also entails great costs. For the EU, for example, multilingualism costs up to € 1.1 billion per year (Gazzola 2016: 35). As not all multilingual institutions can invest great amount of money in translation, some might look for less expensive ways to cope with it. Beyond Koskinen's strategies of translation institutionalization (2011: 59), institutions have started to make use of at least two other strategies to manage translation demand: IT (Information Technology) and non-professional translators, as also commented in Martín Ruano (2014: 4) or in D'Hayer (2012).

Within the Institutional Translation studies, the combination of IT technology and interpreting corpora-based studies have generated a series of important institutional and university studies during the last two decades, like for example the European Parliament Interpreting Corpus (EPIC) and the European Parliament Translation and Interpreting Corpus (EPTIC). Based at the Department of Interpretation and Translation of the University of Bologna, EPIC objective was to *"collect a large quantity of authentic simultaneous interpreting data to produce much needed empirical research on the characteristics of interpreted speeches and to inform and improve training practices"* (Russo et al., 2012: 54). Strictly interconnected with the Interpreting Studies, EPIC involved interpreters, translators, corpus linguists, computational linguists and IT experts who designed and developed a multimedia archive and a corpus of machine-readable transcripts (the EPIC multimedia archive and the EPIC corpus, respectively) (Ibid). For the realization of the EPIC corpus, the European Parliament plenary sittings were recorded off the news channel EbS (Europe by Satellite). Like in the present thesis, all the material thus obtained was digitised and edited by using dedicated software in order to create a multimedia archive of video and audio files. The clips thus obtained were transcribed following specific conventions to create the EPIC corpus. One of the most interesting aspects of this corpus is its complex nature/structure, which allows for carrying out separate searches in the source texts and/or in interpreted texts. In fact, thanks to its inter-modal form, it was possible to contrastively analyse the characteristics of speeches originally delivered in English with those of speeches interpreted into English (comparable corpora), or to compare English source speeches with two interpreted target speeches in Italian and Spanish (parallel corpora). The EPIC video clip archive includes videos of each source language speaker, the audio clips of the corresponding interpreted target speeches, and the transcripts of all the texts. The transcription of audio/video material is indeed of particular relevance to the present thesis. For the transcription methodology, the project adopted a specific convention which has been partially used in the present thesis, as better described in Chapter 3. Apart from studying aspects such as the directionality, the tagging of corpus, lexical density/variety in interpreting (see for example the work of Bendazzoli and Sandrelli, 2005), the EPIC-based studies also focused on disfluency in speech (Russo et al. 2012), which represents an important feature to be examined in the present thesis as well. Regarding the EPTIC (the European Parliament Translation and Interpreting Corpus) corpus (an extension of EPIC), it is interesting to observe that represents a new bidirectional (English<>Italian) corpus of interpreted and translated EU Parliament proceedings. More specifically, it is possible to define EPTIC as an *"intermodal corpus featuring the pseudoparallel outputs of interpreting and translation processes, aligned to each other and to the corresponding source texts"* (Bernardini et al., 2016: 1). In relation to the present thesis, the work by Bernardini et al. has shown that *"interpreted texts are simpler than translated ones and that mediated texts are simpler than non-mediated ones in both English and Italian"* (ibid: 20).

To complete this overview on Institutional Translation studies, it is also important to describe the debate on the role of **English as lingua franca** in the European Union's institutions and in the international organizations, in general. In this respect, some preliminary considerations are to be made regarding the use of English within the international organizations. During the creation of the present study's database, an important challenge will be to obtain a representative sample of English language varieties across the international production/publishing of speeches on climate change. From a theoretical perspective, reference was partially made to Kachru's (1985) *"Three Circles of English"* model and, above all, to Modiano's model (2017). According to the former model, the spread of English developed across the world terms of three concentric circles: the Inner Circle, the Outer Circle and the Expanding Circle (see Figure 12 below).

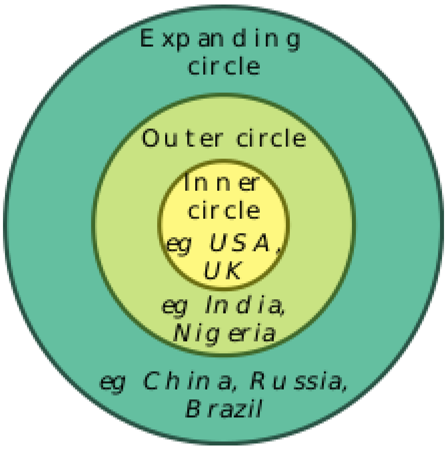


Figure 12 - Kachru's "Three Circles of English".

Each circle represents *"the type of spread, the patterns of acquisition and the functional domains in which English is used across cultures and languages"* (Kachru, 1985: 12). As described by White, "*the Inner Circle refers to the traditional bases of English, dominated by the mother-tongue varieties, where English acts as a first language (White, 1997)"*. These countries are the U.S., the UK, Canada, Australia and New Zealand. More specifically, to continue with the description of the model:

> *"The Outer Circle consists of the earlier phases of the spread of English in non-native settings, where the language has become part of a country's main governmental*

> *institutions, and it plays an important 'second language' role in a multilingual setting. Most of the countries included in the Outer Circle are former colonies of the UK or the U.S., such as Malta, Malaysia, Singapore, India, Ghana, Kenya and others. (Rajadurai, 2005)".*

In the scheme it is finally possible to find the countries where English is learnt as a foreign language; these countries are not territories of former colonization of the UK or the U.S., but they use English as the most useful vehicle of international communication (White, 1997). They represent the so-called "Expanding Circle" and include countries like China, Japan, Greece and Poland. However, being Kachru's model largely criticised in the scientific literature, a better perspective is probably offered by the works by Jenkins (2009), in which the role of English as lingua franca (ELF) is pointed out within the European Union's institutions and other international organizations like F.A.O. of the United Nations, for example. The centripetal model by Modiano (2017) is also considered in the present study because it accounts for other important aspect of English use: (i) English in a post-Brexit European Union and the politics of language within the EU; (ii) the genesis of 'second-language varieties' of English within the European context; (iii) the status of English in European education; and (iv) the development of so-called Euro-English.

If on the one hand, as already mentioned, the protection of linguistic diversity and multilingualism in Europe is crucial, many scholars argue that *"there is also a need for a common language of communication to which the majority of Europeans have access"* (Cogo and Jenkins, 2010: 271). It is sufficiently evident to everyone from the production of written and oral materials in EU that this role is filled by English, since *"it is currently recognised as the most widely used lingua franca within Europe and in many other parts of the world"* (ibid). Within the European Union context, notwithstanding the Brexit transformation of EU institutions, English is still considered as a sort of "working language" but it is not officially regarded as lingua franca. In this respect, Rindler-Schjerve and Vetter point out *"English is not a supranational state language, nor can the lingua franca version of English in the EU be said to carry an exclusively British character"* (2007: 51). Notwithstanding the post-Brexit transformation of the role of English within the EU, the English language is *"shaping itself differently in European contexts from the official languages of the* *two English speaking member states"* (Cogo and Jenkins, 2010: 272). In other words, it is becoming more as a lingua franca (or possibly lingua francas) than as a symbol of national identity. The role of ELF in Europe is perfectly described by the "laissez-faire" attitude towards language policy described by Phillipson (2003).

Over the past two decades, much empirical evidence has been drawn from the analysis of ELF communication. The studies on that material has mostly focused on linguistic features and pragmatic skills underlying such features, but also on the perception of English as a lingua franca. When comparing the native speakers communication with ELF communication, a series of linguistic differences emerge. In terms of phonology/phonetics, for example, the work by Jenkins 2000 should be mentioned, while Peng and Ann (2000) have demonstrated that non-native speakers tend to place stress on the phonetically longest syllable in a word. Other studies have analysed lexical, morphological and lexico-grammatical features of ELF. For example, Pitzl, Breiteneder and Klimpfinger (2008) have pointed that, on the basis of empirical data, ELF speakers create new words and collocations such as 'space time' (where a British English speaker would say 'spare time') and 'severe criminals'. Without describing all studies on ELF features conduced so far, here it is sufficient to comment that research have demonstrated that ELF speakers make frequent and systematic use of certain forms that are not (in some cases, yet) found in native English. To conclude, *"this makes ELF a far more fluid and flexible phenomenon than is understood by the traditional notion of a 'language', and it means that ELF cannot be considered a 'variety' in any traditional sense of the term"* (Cogo and Jenkins, 2010: 278). This consideration will be particularly relevant to the present study when defining the methodology for the Native/Non-Native categorization of speeches included in the present thesis' database.

2.6. Summing up

In this final section of Chapter 2, a critical analysis of the main studies and works reviewed for the purposes of this study is now carried out. First of all, it should be commented that the present literature review is mainly grounded on the theories and studies on Automatic Speech Recognition (ASR) and on the scientific literature within the Accessibility Studies, in addition to examining several works on Neural Machine

Translation (combined with ASR) and on Institutional Translation. Secondly, regarding the previous ASR projects (for example, EU-BRIDGE, TC-STAR and DARPA-GALE), it should be highlighted that they were not based on the combination of ASR with NMT as the technology deployed at that time was not as advanced as it is today. In fact, both ASR and MT technologies were based on now obsolete systems or statistical systems, and they did not include the use of neural networks, in addition to the lack of other important innovations (such as the Cloud technology, the SaaS architecture, the multilingual combination, and the LVCSR requisite). Additionally, it is evident that most of the previous ASR projects examined in §2.2.3.2 did not consider the accessibility of contents or communication for physically impaired people or final users at all, though they were based on, or were sponsored by institutional organizations, universities and international institutions like the European Union. Another weakness of those projects is probably the application of not sufficiently efficacious metrics for the evaluation of accuracy in consideration of accessibility: in most cases, previous projects were based on the WER rate for the ASR output evaluation, and on the BLEU or TER rate for the NMT output evaluation. To sum up, both metrics mentioned above were not adequate, in my opinion, for an effective analysis of accuracy. The BLEU metric is certainly a good measure for the evaluation of accuracy for single segments or sentences but its algorithm does not keep into account the intelligibility and the grammatical correctness of the segments (Castihlo et al., 2018b; Romero-Fresco, 2018). So, for purposes of the evaluation of NMT output in this study, other models or metrics should be considered. In the same way, the WER rate adopted in most of the previous ASR projects is not suitable to examine the accuracy of ASR output for the purposes of accessibility and on a user-informed approach, as it is not based on the evaluation of error seriousness (Romero-Fresco, 2018; Dawson, 2019).

When reviewing the works on Accessibility Studies, it should be commented that the works examined did not sufficiently evaluate the impact of, and the potential benefits offered by ASR on the generation of subtitles for non-hearing people, with only a few exceptions (see for example the works of Romero-Fresco, 2018; Dawson, 2019). Most of Accessibility Studies were mainly focused on the examination of ASR in relation with the respeaking techniques and did not actually offer an assessment of ASR technology as a standalone instrument for the breaking down of barriers in communication, except for a few scholars (for example, Lewis, 2015). The main merit of Accessibility Studies and, in particular, of works on respeaking or subtitles production for media accessibility (like for the example the ILSA or the SMART project) is probably represented by the accuracy evaluation model offered: i.e., a statistical model including the possibility of examining the seriousness of errors and thus the intelligibility of subtitles for non-hearing people. This model will be better examined in Chapter 3 (§3.7). Additionally, it should be remarked that Accessibility Studies contributed to the definition of the criteria and requisites for an efficient methodology of research, which has to be rigorous and transferable (Romero-Fresco, 2020).

With respect to the Interpreting Studies and the analysis of the impact of ASR technology on these studies, it should be commented that they contributed to emphasize the importance and role of ASR in interpreting service, but they mostly focused on the function of query and search-through functionality of this technology for the purposes of the interpreter's work. De facto, these works did not examine the implementation of ASR for the automatic translation/interpretation of speeches, but they preferred to verify the utility of this technology in the booth for rapid information retrieval (automatic translation of acronyms or query through the reference material).

Finally, when considering the importance of studies on Institutional Translation, it should be remarked that the studies mentioned here were certainly efficacious in underlining the role and function of translation/interpretation within the international organizations or institutions but, to the best of my knowledge, they did not evaluate the impact of ASR on enhancing the communication process for accessibility purposes. Yet it should be added that they offered important hints to the present study for the purposes of examining the role of multilingualism and terminology in the communications and speeches held at international organizations or institutions. In fact, the impact of terminological resources on the accuracy of an efficacious ASR + NMT system has never been analysed, to my knowledge, in previous studies with a focus on interlingual and intralingual subtitling for accessibility. After these considerations, the present study will therefore try to propose a methodology of research including an adapted version of existing accuracy evaluation model based on terminology, in a probably innovative way.

3. Methodology

3.1. Introduction

In this chapter, the methodology adopted in the study will be described by firstly introducing the Research Questions (§3.2 below), the database building-up methodology adopted (§3.3 and subsections) with relevant references to the literature. Secondly, an overview of the database inclusion and exclusion criteria as well as its organization will be offered (§3.3.1-2). At this stage, particular attention will be given to the workflow and the possible protocol followed to implement an efficient Automatic Speech Recognition (ASR) workflow (§3.4). Thirdly, the procedures and methods used for transcribing the audio materials will be described (§3.5) by comparing them with other possible methods: an analysis of weaknesses and strengths is presented on this respect. After outlining these preliminary methodological phases, the taxonomy of the various error typologies for the subsequent testing phase is defined, together with the reasons for selecting a statistic quantitative method (§3.6). In particular, for this part of the chapter, a comparison of different models (WER and NER) for the identification of Speech Recognition (ASR) errors and for the evaluation of subtitles accuracy will be added to better identify the most suitable model for the research project (§3.7). An Inter-annotators' Agreement test is also set-up and defined in order to validate the taxonomic scheme here adopted (§3.8). After this step, a presentation and description of the statistical model (NTR) used for the application of Neural Machine Translation is offered to make the measure of accuracy in subtitles for the target language (§3.9) possible. As described below in this chapter, the NTR model (Romero-Fresco and Pöchhacker, 2017) considers the number of words in the audio text (N), the translation errors (T) and the recognition errors (R) to calculate the accuracy rate. Finally, in the Summing up section (§3.10), the potential criticalities of the methodology adopted so far and the possible improvements that could be introduced in further studies will be commented and highlighted.

3.2. Research Questions

This research project focuses on the analysis of English-language output (in the form of subtitles) generated by Automatic Speech Recognition (ASR) and Italian-language output generated by Neural Machine Translation (NMT) technologies (again in the form of subtitles). More specifically, the data input includes official speeches held at international organizations on climate change and its effects on agricultural production (and on related sub-activities). To the best of my knowledge, this kind of discourse has not been investigated so far in the scientific literature as a form of input data for an Automatic Speech Translation (AST) system, including the combined usage of ASR and NMT (see scheme below for a better view of the pipeline in Figure 3.1). The analysed collection is therefore a multimedia database of audio/video materials and their relevant transcriptions in English, subsequently translated into Italian by NMT. Regarding the two written outputs of the present study (the ASR and NMT outputs), it should be clarified that they are examined in the form of subtitles, in an asynchronous workflow. In this respect, it is necessary to add that segmentation of written text was carried out automatically by the ASR software solution implemented. More specifically, the segmentation used is that of VoxSigma and it is organized in time stamps. Finally, it should be explained that, though being analysed in an asynchronous way (one at a time), the ASR and NMT outputs will be considered and examined for the purposes of accessibility (to be reproduced in real time), without taking into account the problem of latency.

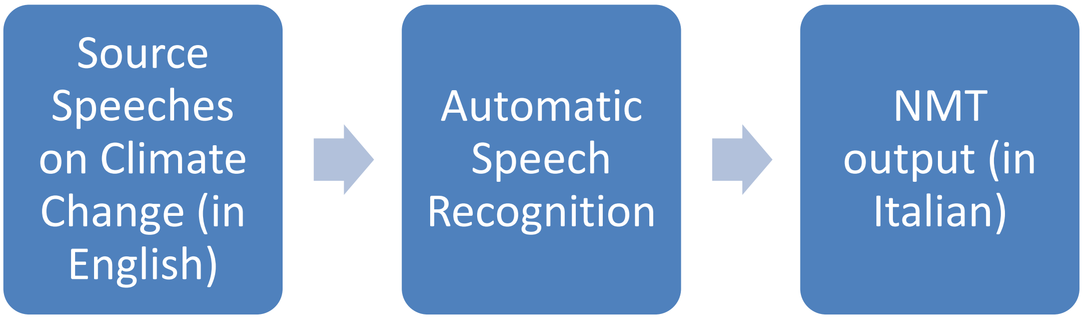

Source Speeches on Climate Change (in English)
Automatic Speech Recognition
NMT output (in Italian)

Figure 3.1 - The present study's pipeline

Despite having some similarities with an interpreting corpus of texts (for example because it aligns files of input and output text generated from the audio/video files) like it was seen in the case of the EPIC project (see §2.5), the present study's database should not be associated, nor compared to an interpreting corpus of texts. In fact, even if ASR and NMT technologies act together like an interpreter in producing an output partially similar to that of human interpreters in an asynchronous workflow, yet the nature and structure of this database is specific. Further considerations in this respect will be presented below in the description of the database. At this stage, before describing the database building phase and the methodology implemented in the study, it is necessary to define the main Research Questions (RQs) at the basis of it.

While in previous studies and projects attention was paid mainly to the usage of ASR technology combined with the intervention of a subtitle editor or respeaker (or in combination with an interpreter), in this study the human mediation role is eliminated by attempting to define a protocol for the usage and setting up of an entire ASR+NMT pipeline as shown in Figure 3.2 below. In addition, it should be highlighted that in previous research projects, a limited or scarce attention was dedicated to the application of domain-specific terminology (*i.e.*, domain-specific termbases or institutionally-approved terminological resources) during the implementation of those technologies for the purposes of accessibility and/or interlingual institutional communication.

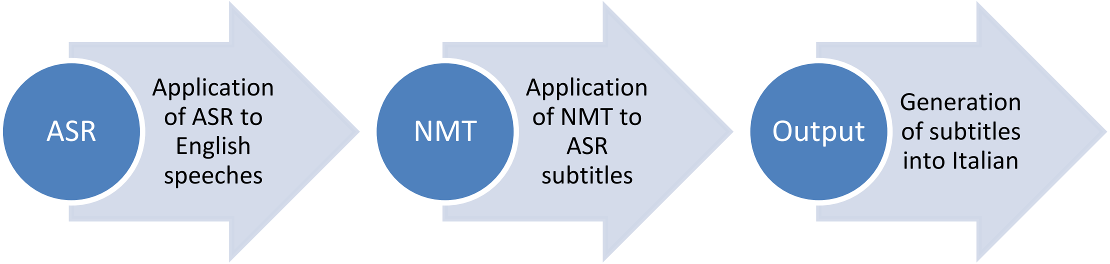

ASR
Application of ASR to English speeches
NMT
Application of NMT to ASR subtitles
Output
Generation of subtitles into Italian

Figure 3.2 – ASR-NMT-based pipeline methodology in this study.

Under the above presented pipeline, it should be clarified that the application of ASR is carried out directly to the English-source speeches, which are transcribed in the subtitle format according to the ASR software segmentation (which is better described below). Subtitles are then processed by NMT to obtain the Italian-target subtitles.

Starting from all these considerations, and taking into account the literature already produced so far in this field of studies, the need for analysing and evaluating the ASR output under a different perspective emerges, including the necessity of better assessing the "significance" of terminological resources in an ASR system and across the entire pipeline shown above. In particular, the main Research Questions for this study are defined as follows.

This study's **Research Questions (RQ)**:

1. Can ASR technology produce accurate output[17] for the breaking down of the barriers of communication in the intralingual context (in the English language)?
2. Can the combination of ASR and NMT provide an accurate output in generating subtitles for the purposes of accessibility in the interlingual context (namely, from English into Italian)?
3. Do domain-specific terminological resources (incorporated into the ASR step of the pipeline) improve the accuracy of interlingual and intralingual subtitles in this study's specific scenario?

These Research Questions are based on the concept of accuracy presented and described in Chapter 2, both for Automatic Speech Recognition and for Neural Machine Translation, especially in the section dedicated to Accessibility Studies (§2.4), where it is evident that the very concept of accuracy is interconnected with quantitative and qualitative measures and to the use of statistical models, which are described in detail in this chapter. The present study will therefore make a choice in this respect for a specific-context definition of accuracy. These research questions certainly pose a series of challenges and criticalities. In particular, the evaluation of an appropriate tool set of ASR and NMT technologies is to be carried out in order to identify available software solutions, as well as defining the technical features and

[17] The definition of accuracy will be given later on this chapter.

specifications required for the purposes of audio/video file processing, and the software utilities that are necessary to generate subtitles. Secondly, the identification of a protocol and reference industry standards are also specified to define what is meant by "accurate output" in the RQs above. In this respect, the taxonomy of errors for the testing phase will be defined (see §3.6 below) to better specify the various error typologies and, possibly, a relevant degree of error grading. This will contribute to establishing the accuracy of final output according to the industry's minimum accuracy requisite. More specifically, accuracy will be here based on the statistical measures adopted in previous studies (WER, NER models) on Automatic Speech Recognition (see, for example, Eugeni, 2008; Romero-Fresco, 2016) and on Neural Machine Translation (see, for example, Dawson, 2019), within a user-focused approach oriented towards accessibility for non-hearing people (as seen in Chapter 2, §2.4). Furthermore, in order to respond to the questions above, a quantitative, statistical approach will be adopted to try to make a general evaluation of Speech to Text output (*i.e.*, real-time intralingual and interlingual subtitles) generated by ASR and NMT software, with the objective of assessing its accuracy (Romero-Fresco, 2011: 104), where the concept of accuracy is simply connected with the NER/WER rate achieved (see §3.6 and 3.7 below).

Before continuing with the presentation of the present study's methodology, it is necessary to define the discourse context and the database of audio/video materials used here.

3.3. The Database

This section describes the corpus-based studies referenced to in the creation of this study's database, how the database was collected and prepared, and it gives a description of the database itself. The decision of selecting a database format in place of a corpus is based on the considerations that the present study includes a collection of audio/video files, as well as an archive of automatically generated transcriptions (in the subtitles format) and of the corresponding gold standard transcriptions created by the author of this thesis. Hence the materials are not produced by multiple authors like in a corpus (for example, in EPIC, the authors include translators, interpreters, speakers etc.) and the alignment of contents (between audio/video files and the relevant transcriptions) is not carried out.

The complete database of this study is available in **Appendix A** for further consultation. Starting from the assumption that no unanimous definition is provided for a database, by using the general definition used in Computer Sciences, it is possible to generally maintain that a database is an organized collection of data, generally stored and accessed electronically from a computer system. To use the definition offered in the Encyclopaedia Britannica, a database or base of data can be described as:

> *"any collection of data, or information, that is specially organized for rapid search and retrieval by a computer. Databases are structured to facilitate the storage, retrieval, modification, and deletion of data in conjunction with various data-processing operations". (Britannica, 2020)*

In computer science and multimedia studies, in general, a database is stored as a file or a set of files. The data and information in these files may be broken down into records, and each of these records consists of one or more fields. Fields are the basic units of data storage, and *"each field typically contains information pertaining to one aspect or attribute of the entity described by the database"* (Britannica, 2020). Although the term database is widely applied to any collection of information in computer files, a database in the strict sense provides cross-referencing capabilities, that is to say the possibility of carrying out queries across the different records and files. The present study's database is a collection of naturally occurring samples of texts in the electronic format, and it was constructed according to a number of coherent selection criteria, including the **authenticity** of texts and their **representativeness**. The database so realized includes material in the electronic format and it incorporates authentic material. As it happens with an institutional interpreting corpus (which is a different concept from computational linguistics) like in EPIC, the source materials (*i.e.*, audio and video materials) incorporated into the database are all authentic. Furthermore, as specified by McEnery and Wilson (1996: 87) for a corpus of text, also in the case of this study's database it is necessary to comply with the representativeness requirement as *"a body of text which is carefully sampled to be maximally representative of a language or language variety"*.

Starting from these general assumptions (authenticity, and representativeness), the database compiled for this study is composed of four main components: *i.e.*, the source audio/video files available in the English language, the relevant file transcriptions (again in the English language) produced both manually and by the usage of ASR technology (namely, through Google Speech Recognition via YouTube/Descript application and through *VoxSigma* by Vocapia Research), and, finally, the automatic translations carried out through Neural Machine Translation technology (in the Italian language). See Figure 3.3 below for a better understanding of the database components.

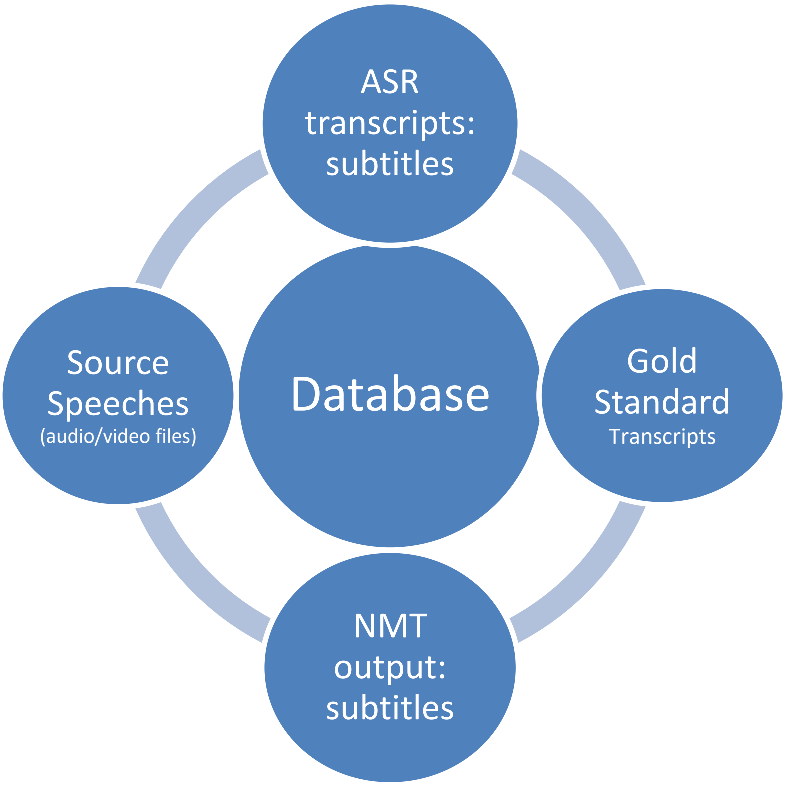

ASR transcripts: subtitles
Source Speeches (audio/video files)
Database
Gold Standard Transcripts
NMT output: subtitles

Figure 3.3 - Components of the present study's database.

The typology of text derived from the source speeches consists in official speeches (by mono-speakers) held by officials, politicians or institutional spokespersons at conferences, summits, committee or institutional sittings on the topic of climate change and its effects on agricultural production (and on related sub-activities). Whether held as *impromptu* or read-out speeches, the oral source speeches therefore represent the "core" element of this study database and, in this respect, it should be underlined that speech, particularly if *impromptu*, is among *"the most difficult and expensive [language varieties] to acquire, difficult to classify and manage"* in relation to the creation and analysis of spoken corpora, as commented by Sinclair, 1996 (in Bernardini et al., 2018: 22). All *impromptu* speech features will be reviewed and discussed in an in-depth manner in the next chapter, dedicated to the analysis of data.

To further describe this study database, it is possible to maintain that, collectively, the four components of the database (see Figure 3.3 above) constitute a comparable, searchable set of speeches on climate change. More precisely, the different events (i.e., speeches) are comparable to one another because they were selected according to specific inclusion criteria (described in §3.3.1), but the four components of the database are de facto different versions of the same single event (audio/video content, ASR output subtitles, NMT output subtitles). In addition, the ASR and NMT outputs can be labelled as *"comparable"* because all components gather similar samples of texts (Tognini-Bonelli, 2001: 7): namely, they include the same segment units, and they refer to the same initial event. In addition, part of the database is "searchable" (namely, the reference transcriptions, the ASR output transcriptions and the NMT output) because it allows executing searches of text parts or words in the electronic format (the searching operation can actually be carried out in the gold standard transcriptions, in the ASR and NMT outputs).

3.3.1. The Database requisites

In the selection of this study's source audio/video files and in the building up of the database, a series of requisites were identified in order to have a comparable set of texts and also to satisfy the principle of representativeness. These requisites are as described in the list below.

- **Authenticity:** all texts in the database are naturally occurring instances of communications, *i.e.*, authentic oral speeches held within a restricted selection of international organizations, namely, the Food and Agricultural Organization (F.A.O.) of the United Nations, or speeches held at summits, debates, committee or plenary sessions of the United Nations and of the European Parliament, before an international audience of experts and non-experts; all audio/video contents have also been published on the official websites or channels of those organizations[18].

[18] The official websites and channels are detailed in §3.3.2 below.

- **Comparable institutional settings:** all the hosting institutions are international organizations (both governmental and non-governmental).
- **Topic and timespan:** consistency is maintained in terms of topic (oral speeches on climate change), and timespan (all audio/video contents were produced and published between 2013 and 2019).
- **Single Speaker:** speeches held by multi-speakers are not considered in this study; all speeches in the database are mono-speaker-based, English-language Native or Non-Native speakers and cover a similar institutional function/role, *i.e.*, they are Members of Parliament, or high officials/charges at the international organizations selected for the study.
- **Audio quality:** the audio material collected and analysed for the study is consistent in terms of audio quality (in other words, all parts of the speech are clearly audible, with no interruptions) and in optimal audio condition (clarity). In the present study, noise-disturbed audio/video material is excluded (i.e., files with background road traffic, background voices, music or other sounds covering most of the source content). Yet some reduced portions (a few seconds) in a little partition of sample files may include noise or other sources of disturbance (applauses, laughter, etc.).
- **Specialization of texts:** all contents are related to climate change and its effects on agricultural production, fishery, farming, and other human economic or production activities. The terminology is considered as specialized in all audio/video files.

As mentioned above, the setting up and composition of this database may share a few similarities with corpora of texts and, in particular, with interpreting corpora where a gold standard is often used for an evaluation of quality (see D'Hayer, 2012; Fantinuoli, 2018). Like it happens in an interpreting corpus-based study (for example, in EPTIC-based studies), it is possible to ascertain that the present study *"is always based on a comparison between corpora of different types so that, in translation studies, a corpus is actually always a combination of at least two subcorpora"* (Zanettin 2013: 26). In particular, both sets of data (interpreting corpora and this database) are indeed based on sub-databases relating to speech material, they include an audio input (the source speeches) and the interpreting corpus in the target language: in the case of the present database, this subset of data is represented by the ASR+NMT output.. Additionally, in the analysis of these sub-databases, there is always a translation of input from one language to another, or a conversion of signs/inputs into an accessible format. More precisely, in common with the corpus-based interpreting studies, in the realization of this database it is possible to identify features like the **oral communication** (speech), the **translation** (in this project represented by ASR and NMT), and the **multi-modal nature** of the database. The database does in fact include different modality components, *i.e.* sub-databases which bring together different modes (audio/video material, transcriptions of oral communications, NMT output in Italian). Yet it should be remarked that, in general terms, the nature of the present database cannot be compared with that of interpreting corpora under many aspects. For example, if the database created here is compared to the European Parliament Translation and Interpreting Corpus (EPTIC)[19], it is possible to point out, first of all, that no human interpreters are involved in the present study. Secondly, the interlingual translations are generated by a NMT solution and not by the institutions where the speeches were held.

3.3.2. The Database organization

After reviewing the general criteria at the basis of the study's database definition and the requisites to be satisfied for its creation, the procedure followed in the audio/video file collection and the database compilation is now described in detail. As mentioned above, all audio/video files are official speeches on climate change given at the Food and Agriculture Organization (F.A.O.)[20], the European Parliament[21] or the United Nations[22], including conferences hosted by these organizations. In particular, all audio/video files are made publicly available on their official Websites or official channels (namely, on *YouTube* platform) for anyone willing to listen to or watch them. In the case of the European Parliament (EP)'s speeches, the EP portal[23] was also consulted. All these multimedia contents are therefore free and do not require any registration or login to the organizations' Web pages or channels in order to be consulted. For the purposes of this study, there was not therefore any need to ask for a

[19] EPTIC: available for consultation on https://corpora.dipintra.it/eptic/

[20] Food and Agriculture Organization (F.A.O.) channel on YouTube: https://www.youtube.com/user/FAOoftheUN

[21] European Parliament channel on YouTube: https://www.youtube.com/EuropeanParliament

[22] United Nations' channel on YouTube: https://www.youtube.com/unitednations

[23] European Parliament's official portal: https://www.europarl.europa.eu/portal/en

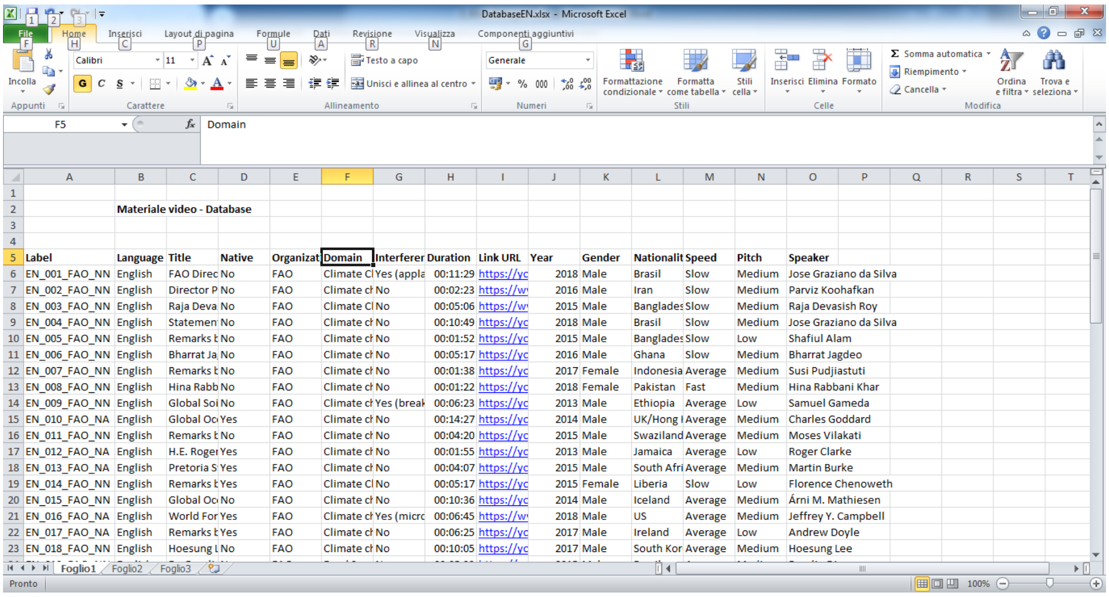
special authorization to use the audio/video materials in question, as the contents are public.

The database construction phase involved a data collection and review phase, during which it was possible to search for and identify a set of speeches responding to the requisites indicated before. During that phase, the main difficulty was represented by the identification of speeches held by single speakers, thus eliminating all video/audio contents with two or with multiple speakers (also excluding all speeches in an interview format), also to respond to the consistency requisite defined above. As one of the main requisites is the authenticity of contents, it was also difficult to find a vast set of institutionally-approved or published material on the official channels of the organizations analysed here. The inclusion criteria of audio quality, single speaker and the necessity of collecting an almost equivalent number of speeches held by Native and Non-Native speakers generated a certain difficulty. A further challenge was represented by the fact that most of the materials published on the Web consist of debates, with multiple speaker voices that can bias results due to the overlapping of discourse. For example, a lot of the material initially collected (but not included) was from the national governments' debates or parliamentary sessions: *e.g.*, the UK Government's Question Time, or the EU Parliament's plenary sessions debates. All these video materials were therefore excluded from the database.

The database built-up during this phase of the research project was described on a Microsoft Excel spreadsheet and it was organized into 15 columns (see Figure 3.4 below for a screenshot of the Excel sheet: the complete database is consultable in Appendix A), each including a specific piece of information.

To describe how this database is organized, starting from the left, in **Column A**, the database indicates the label which is associated to each audio/video file in order to efficiently manage and organize the data collection, as well as for an easier reference during the analysis and comparison of results (see Chapter 4). This label was defined in a simple way, reporting only the language category (EN = English), a 3-digit number identifying the different files (between underscores "_"), followed by the conventional abbreviation of the name of the international organization hosting or publishing the speech (*e.g.*, FAO), and, finally, by the English variety indication: *i.e.*, Native (NA) or Non-Native (NN). For further details on the English variety representation, see the description of Column D below. **Column B** indicates the source language (*i.e.*, the

English language); **Column C** reports the titles of the speech or video content as assigned or published on the official Website or channel.



Materiale video - Database

Label	Language	Title	Native	Organizat	Domain	Interferer	Duration	Link URL	Year	Gender	Nationalit	Speed	Pitch	Speaker
EN_001_FAO_NN	English	FAO Direc	No	FAO	Climate Cl	Yes (appla	00:11:29	https://yc	2018	Male	Brasil	Slow	Medium	Jose Graziano da Silva
EN_002_FAO_NN	English	Director P	No	FAO	Climate ch	No	00:02:23	https://w	2016	Male	Iran	Slow	Medium	Parviz Koohafkan
EN_003_FAO_NN	English	Raja Deva	No	FAO	Climate Cl	No	00:05:06	https://w	2015	Male	Banglades	Slow	Medium	Raja Devasish Roy
EN_004_FAO_NN	English	Statemen	No	FAO	Climate ch	No	00:10:49	https://yc	2018	Male	Brasil	Slow	Medium	Jose Graziano da Silva
EN_005_FAO_NN	English	Remarks b	No	FAO	Climate ch	No	00:01:52	https://yc	2015	Male	Banglades	Slow	Low	Shafiul Alam
EN_006_FAO_NN	English	Bharrat Ja	No	FAO	Climate ch	No	00:05:17	https://yc	2016	Male	Ghana	Slow	Medium	Bharrat Jagdeo
EN_007_FAO_NN	English	Remarks b	No	FAO	Climate ch	No	00:01:38	https://yc	2017	Female	Indonesia	Average	Medium	Susi Pudjiastuti
EN_008_FAO_NN	English	Hina Rabb	No	FAO	Climate ch	No	00:01:22	https://yc	2018	Female	Pakistan	Fast	Medium	Hina Rabbani Khar
EN_009_FAO_NN	English	Global Soi	No	FAO	Climate ch	Yes (break	00:06:23	https://yc	2013	Male	Ethiopia	Average	Low	Samuel Gameda
EN_010_FAO_NA	English	Global Oc	Yes	FAO	Climate ch	No	00:14:27	https://yc	2014	Male	UK/Hong	Average	Medium	Charles Goddard
EN_011_FAO_NN	English	Remarks b	No	FAO	Climate ch	No	00:04:20	https://yc	2015	Male	Swaziland	Average	Medium	Moses Vilakati
EN_012_FAO_NA	English	H.E. Roger	Yes	FAO	Climate ch	No	00:01:55	https://yc	2013	Male	Jamaica	Average	Low	Roger Clarke
EN_013_FAO_NA	English	Pretoria S	Yes	FAO	Climate ch	No	00:04:07	https://yc	2015	Male	South Afri	Average	Medium	Martin Burke
EN_014_FAO_NN	English	Remarks b	Yes	FAO	Climate Cl	No	00:05:17	https://yc	2015	Female	Liberia	Slow	Low	Florence Chenoweth
EN_015_FAO_NN	English	Global Oc	No	FAO	Climate ch	No	00:10:36	https://yc	2014	Male	Iceland	Average	Medium	Árni M. Mathiesen
EN_016_FAO_NA	English	World For	Yes	FAO	Climate ch	Yes (micro	00:06:45	https://w	2018	Male	US	Average	Medium	Jeffrey Y. Campbell
EN_017_FAO_NA	English	Remarks b	Yes	FAO	Climate ch	No	00:06:25	https://yc	2017	Male	Ireland	Average	Low	Andrew Doyle
EN_018_FAO_NN	English	Hoesung L	No	FAO	Climate ch	No	00:10:05	https://yc	2017	Male	South Kor	Average	Medium	Hoesung Lee

Figure 3.4 – Screenshot of the Database spreadsheet.

For the definition of the Native/Non-Native variable in **Column D**, the considerations already made in §2.5 implied the categorization of English use in the database speeches according to two categories only for simplification: Native and Non-Native. While every effort was made in the present thesis to represent as many varieties of English as possible, Modiano's simplified model was chosen as the main reference scheme: Native contents in the database are those audio/video files where the speaker belongs to the Inner Circle (namely, the United Kingdom, Australia, Ireland, New Zealand, the United States, Canada), while Non-Native files are the remaining situations (i.e., speakers largely belonging to the Outer Circle/Expanding Circle or where English is used as *lingua franca*). This procedural decision allows for a simpler accuracy assessment of the ASR output, also considered the fact that the acoustic models of marketed ASR technology (see §3.4.1 below) are mainly designed on L1 variety of English (*i.e.*, the Inner Circle). It is also important to underline that, when English pronunciation is analysed, the correct version considered here is that implemented by the ASR: i.e., the standard English variety of the United Kingdom and United States.

Next to Column D, the abbreviation conventionally used for the name of the hosting international institution is indicated (*e.g.*, FAO or EP) in **Column E**. In

Column F, the main domain and the specific sub-domain of the speech are indicated. However, it should be added that the entire database of audio/video files includes specialized contents in the field of climate change and agriculture. This column is thus created to specify the potential subdomain of each single file. The indication was assigned by taking into consideration the topic title of the speech, and the topic or title of the conference/event into which that speech was given.

As far as **Column G** is concerned, it is interesting to point out that this piece of information indicates the presence (or absence) of interference noises or other sources of background noise, possibly compromising or disturbing the quality of the audio signal. In particular, noises or background noise such as music, applauses or the echo effect produced by the microphone, including background road traffic, are indicated. These indications may be of particular interest for the analysis of disfluency in ASR and other elements of speech in the analysis phase. Yet it should be underlined that these noises or background events are only interfering with the speech for a few seconds (as seen in the requisite of Quality defined above in the setting-up of the database). In fact, audio/video files with long-interference noise were excluded from the database.

Continuing with the description of the database organization, in **Column H** it is possible to find information on the duration of the audio/video file, while in **Column I** the URL address of the relevant file published on the official Website or channel is indicated. In **Column J**, the year of publication is reported, while in the following **Column K** the gender (male or female) for the speaker is specified. The nationality of the speaker is specified in **Column L**, adding a further level of specification with respect to the speaker's origin and English-language variety.

For **Column M** (indicating the Speech Speed), a more detailed consideration is required to establish how to measure the speech rate of the speaker. As far as the speech rate is concerned, the speech (or fluency) rate is the speed at which a person speaks, measured in words per minute (wpm). More in detail, in Speech Recognition technology, as indicated on SpeakerHub (2017) website, speech rate is considered as follows:

- ***Slow****: with less than 110 wpm.*
- ***Average*** *(or Conversational): between 120 wpm and 150 wpm.*
- ***Fast:*** *more than 160 wpm.*

Related to the speech rate is also the speaker's tone or pitch. This value (specified in **Column N**, see the database in **Appendix A**) is measured by using a simple free-downloadable tool developed by P. Boersma and D. Weenink from the Department of Phonetic Sciences, at the University of Amsterdam: *PRAAT*[24]. This tool allows creating a spectrogram of each audio file, measuring the pitch value among many others: *e.g.*, the intensity of voice; see Figure 3.5 below for an example of spectrogram. For a male speaker the average pitch range is normally 85 to 180 Hz, while for female speakers lies between 165 to 255 Hz (or even higher). In the present study, values below 100 Hz are considered low, between 100-120 Hz are medium and above 120 Hz are high (for male speakers). For women values below 180 Hz are low, between 180-200 Hz are medium, and above 200 Hz are high.

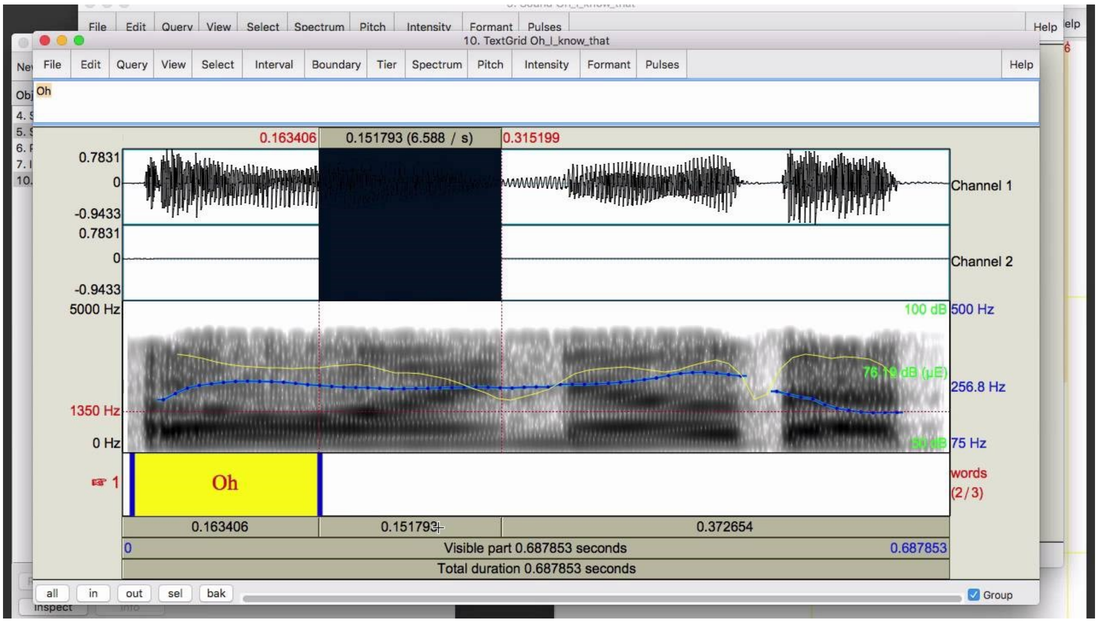


Figure 3.5 – Example of pitch measurement (spectrogram) with PRAAT.

Both the speech rate and the pitch value were useful in the testing and analysis phase described in chapter 4. To conclude with the description of the database, it should be added that in the last **Column O**, the name of the speaker is reported for further reference in the following analysis of the present study.

[24] Downloaded from the website of the University of Amsterdam: http://www.fon.hum.uva.nl/praat/download_win.html

3.4. Database Building and Processing Workflow

During the database building phase, it was necessary to consider previous projects on Automatic Speech Recognition applications and, in particular, select appropriate ASR solutions and set up a processing workflow capable of elaborating the source materials in an asynchronous sequence. This required dividing the work into two steps: firstly, 1. the selection of ASR technology, and, secondly, 2. the definition and configuration of a protocol to be followed in the automatic and manual processing of data in an asynchronous sequence.

3.4.1. Selection of ASR Technology

As already seen in §2.2.2 and §2.2.3 (and its subsections), in order to obtain higher accuracy in the output (subtitles) generated by the ASR component of the pipeline examined here, the ASR technology implemented in the present study has to comply with specific requirements. For the purposes of selecting the ASR technology, the most important criteria for an efficient system were reviewed on the basis of the indications provided by the industry and by scholars from similar studies on ASR. This led to the following considerations. While meeting the requisites of easy-to-use interface, the minimal computer requirements, the multilingual acoustic model (English and Italian languages are available), and the LVCSR, Dragon Naturally Speaking by Nuance was excluded because the solution is speaker dependent, and it is thus incompatible with the audio/video material used in the present study. Even if this software can be "trained" to the speaker's voice, yet it does not allow processing audio files from different speakers on an immediate basis. Additionally, the software was excluded for not providing for the convenient functionality of a SaaS service (based on the cloud) and for not meeting the Augmented Terminology requisite. The other second software reviewed in the early phase, i.e., Microsoft Skype Translator, was found to offer all the main features integrated into GSR (as described above), but its use requires the purchasing of a subscription for the processing of a large number of files. Its functionality (even for the basic features) is conditioned to the payment of a fee for a high volume of files like in the present study. The review of ASR technology thus opted for the selection of VoxSigma software and GSR engine through YouTube and Descript platforms. In fact, even if their use provides for the payment of a reduced fee (in the case of VoxSigma) or for a free registration (in the case of YouTube and Descript), both solutions can cope with the processing of a high volume of files, including the advanced features required by the present study: i.e., Augmented Terminology, cloud-based functionality and learning capability (the "trainable" requisite).

3.4.2. File Processing Workflow

After evaluating and selecting the most appropriate ASR technologies for the purposes of this study, it was necessary to concentrate efforts on configuring and setting up a general protocol for the data processing workflow (see the workflow in Figure 3.6 below).

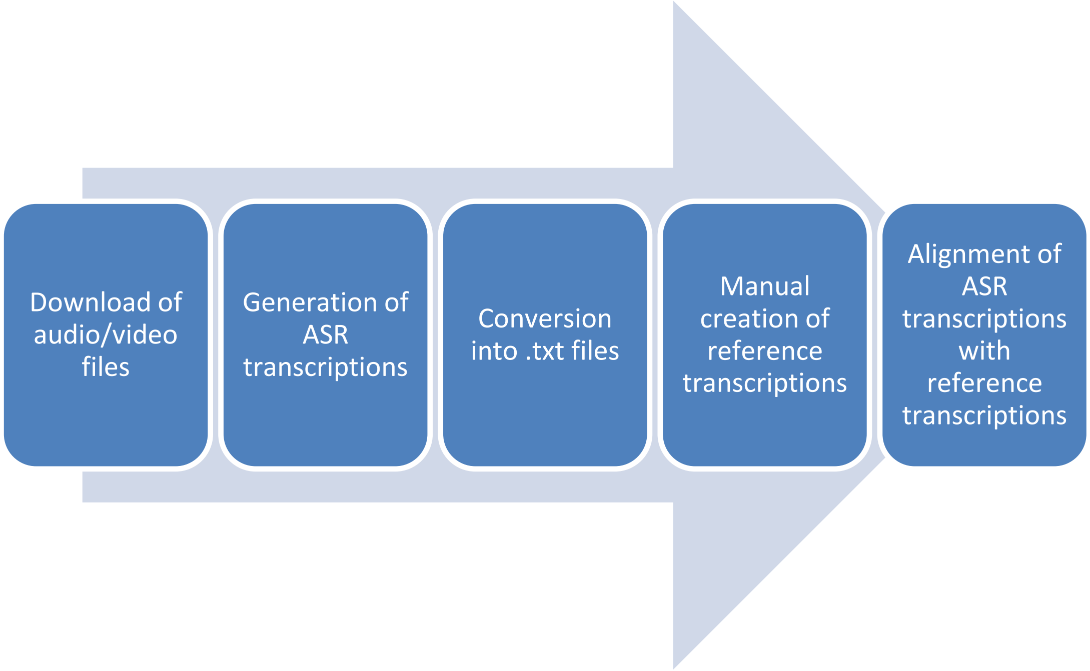

Download of audio/video files
Generation of ASR transcriptions
Conversion into .txt files
Manual creation of reference transcriptions
Alignment of ASR transcriptions with reference transcriptions

Figure 3.6 – General workflow for data processing.

Looking at the Figure above, it is possible to observe that the workflow was organized into 5 steps, which can be described as follows.

Step 1. - Download of audio/video files: all speech files were downloaded on the PC locally in the *.avi* or *.mp4* format by clicking on the URL specified on the Database sheet previously prepared.

Step 2. – Generation of ASR transcriptions: in this step, two different approaches were adopted on the basis of the software used. For GSR engine (via YouTube), the strategy for obtaining the automatic transcriptions was that of using a simple, free utility denominated *DownSub*[25]. This operation included the copying of the URL for the YouTube video link on the utility Website directly. From here, it was then possible to obtain the transcription file in the subtitle format (*.srt*) with time indexing (and the possibility of pairing it to the video images). For all audio/video files not available on *YouTube* (and thus not available for the automatic transcription) portal, the Descript app was used[26]**,** allowing executing the automatic transcription of files through GSR engine. When using VoxSigma, a prior conversion of video files into audio files was required in order to process the automatic transcriptions. This operation was simply done by using free, open-source audio software *Audacity*[27] (see Figure 3.7 below for a screenshot of Audacity). The generation of ASR transcriptions was subsequently completed automatically by using VoxSigma processing command, as in Figure 3.8 below.

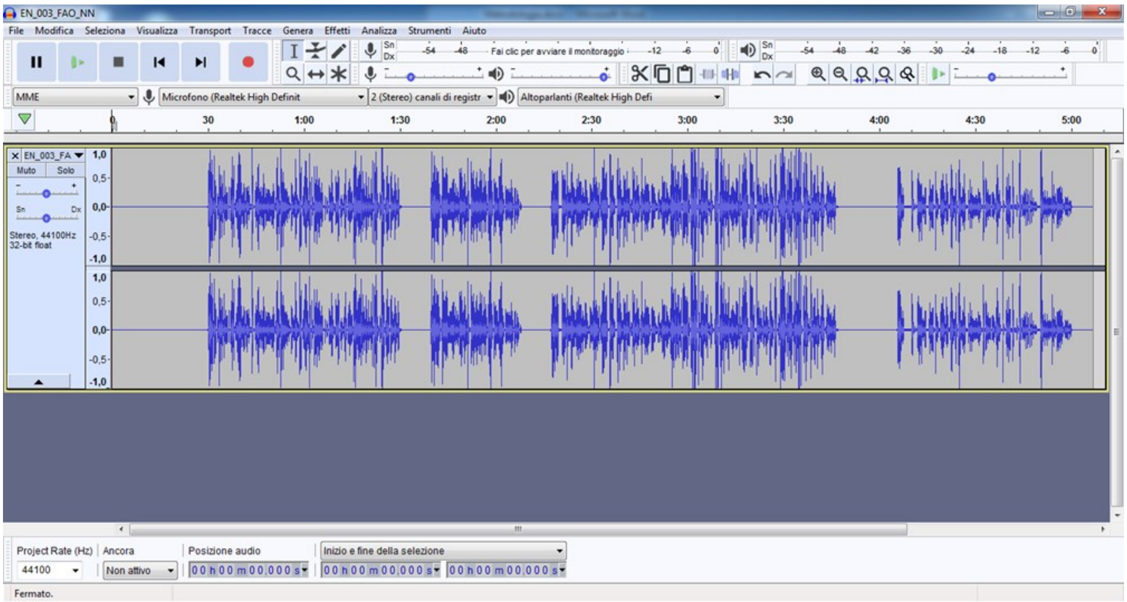
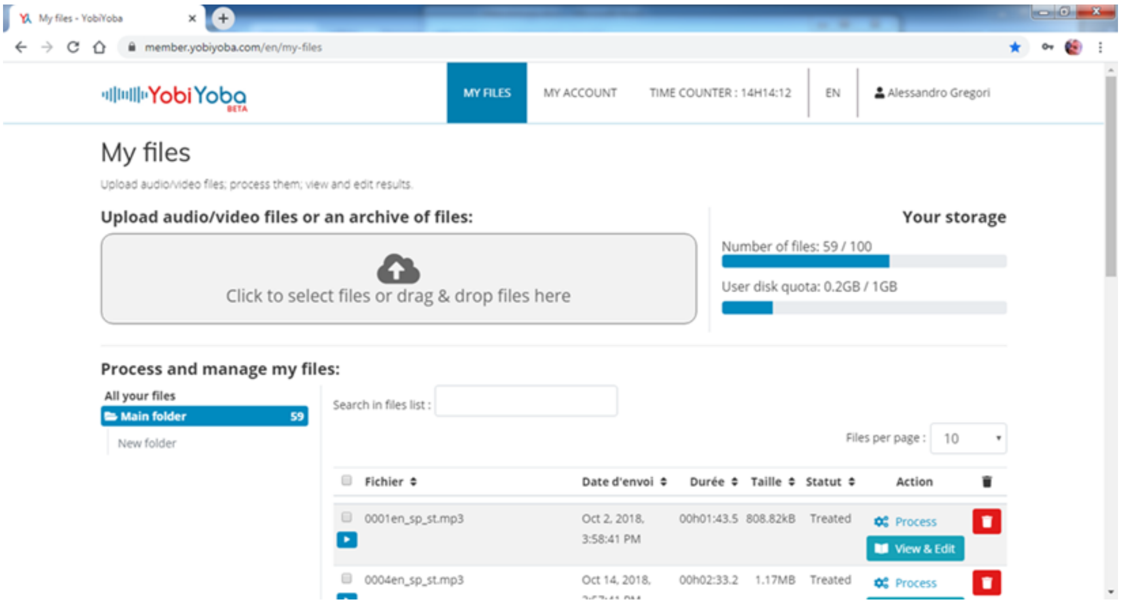


Figure 3.6 – Audacity software used for the conversion of video files into audio files.

[25] Link for download and further information: https://downsub.com/
[26] Link for download and further information: https://www.descript.com/
[27] For download and further information: https://www.audacityteam.org/



My files

Upload audio/video files; process them; view and edit results.

Upload audio/video files or an archive of files:

Your storage

Number of files: 59 / 100

User disk quota: 0.2GB / 1GB

Process and manage my files:

All your files
Main folder 59
New folder

Search in files list :

Files per page : 10

Fichier	Date d'envoi	Durée	Taille	Statut	Action
0001en_sp_st.mp3	Oct 2, 2018, 3:58:41 PM	00h01:43.5	808.82kB	Treated	Process View & Edit
0004en_sp_st.mp3	Oct 14, 2018,	00h02:33.2	1.17MB	Treated	Process

Figure 3.8 – Screenshot of VoxSigma transcription processing

Step 3. – Conversion into .txt files (with no tag): in the case of VoxSigma and Descript app, this operation was simply carried out by using a special Export command. On the contrary, when using YouTube application (GSR engine), this step was made possible by installing and making use of a free software application denominated *SubtitleEdit.exe*[28]. It allowed exporting the *.srt* file created in the previous step, eliminating all unnecessary tags and time indexes. Among the various *Export* options, the user can select *Remove styling* to remove all formatting. The result is a clean, unformatted *.txt* file (no punctuation is provided). As far as the time spent by the solution to process each file is concerned, it should be underlined that the audio file processing was carried out on remote (both ASR technologies are cloud-based), so the time required for completing the process depended on Internet speed connection, and on the platform server's workload intensity in that given moment, apart from the file size. On average, it was observed that each file took about 6-10 minutes to be processed by the software (both in the case of *Descript* and of *VoxSigma*: for files with about 10-minute duration).

Step 4. – Creation of the reference transcriptions (gold standard): for convenience, it was decided to create the transcriptions starting from VoxSigma's ASR output, as it provided for a better organized text with time stamps. In the following section on Transcription and annotation, a definition of time stamp will be provided, together with the criteria adopted for the segmentation and organization of subtitles

[28] Link for download and further information: https://www.nikse.dk/subtitleedit

and the relevant text. At this stage, it is sufficient to underline that this part of the workflow involved much of the efforts required for the database building-up phase. In fact, it was necessary to listen to all the video/audio files collected, and to manually carry out the transcription of the speeches in a *.txt* format file. Considering the often not so clear pronunciation of words, this task was particularly difficult and time-consuming because almost half of the files are from Non-Native speakers. The second reason for this is represented by the fact that the transcription work was made more difficult by speaker speech rate (often rapid in the case of Native speakers). On average, it was estimated that a file of about 10 minutes required, approximately, a 1-hour manual transcription work, also considered the proof-check work carried out at the end of it.

Step 5 – Alignment of ASR transcriptions with reference transcriptions: this operation was also very consuming in terms of time and efforts as the alignment of VoxSigma's and GSR's transcriptions with the gold standard transcriptions was carried out manually, on the basis of the automatic time stamp organization generated by VoxSigma. The alignment of texts was produced in Excel spreadsheets. Excel files allowed for the insertion of transcriptions and annotation data in practical way.

3.5. Transcription and annotation

After describing the database building process in detail and the workflow for the processing of the automatic transcriptions by the software solutions selected here, it is now necessary to define and specify the criteria adopted in the manual processing of the reference transcriptions (the study's Gold Standard) for the 55 video/audio files included in the database. As maintained by Thompson (2005) (in Russo et al., 2012: 57), one of the fundamental steps in the creation of a spoken corpus of texts is indeed transcription. As already mentioned in §3.4, the manual reference transcriptions of all speeches (to be used in the subsequent analysis of the final outputs) were carried out starting from the automatic transcriptions generated by the software *VoxSigma* rather than making a "transcription from scratch". The rationale of this decision is also based on the assumption that, given the automatic nature of the workflow (like in a real time situation), the modification of the segmentation in units of meaning would be a sort of human intervention. The default automatic segmentation indeed offered a valuable basis for quickly creating the final reference transcriptions (or gold standard transcriptions), since it speeded up the process of manual transcription. However, it should be specified that there was no double checking for the transcribing work conducted by the author of this thesis. This approach is similar to the method adopted in the EPIC research project, where transcriptions were made starting from the official European Parliament's verbatim reports (see Russo et al., 2012: 58). The method also offered the advantage of a predefined segmentation based on machine-generated time stamps (in this case, VoxSigma's time stamps), allowing for an easier comparative analysis in the subsequent phase of this study. More specifically, time stamping "*refers to the process of adding timing markers – also known as time-stamps – to a transcription"* (JBI Studios' Blog, 2017). The time-stamps (also denominated as "time offset values") can be added at regular intervals, or when certain events happen in the audio or video file. As explained in Google Cloud Speech to Text (2020) website: *"time offset values show the beginning and end of each spoken word that is recognized in the supplied audio. A time offset value represents the amount of time that has elapsed from the beginning of the audio, in increments of 100ms"* (Google Cloud Speech-to-Text, 2020).

It should also be mentioned that, as for the previous phases of the database building-up process, a certain balance between practicality and representation of speech features was kept during the transcribing phase. On the one hand, it is almost impossible to reproduce all the characteristics of speech in writing as there are several levels of communications (*i.e.*, linguistic, prosodic and extra-linguistic), and each level comprises a multitude of features (as also mentioned in Russo et al., 2012: 57), for example, pauses, repetitions, hesitations, or background noise. On the other hand, the study adopted a series of guiding principles as inspired by best practices and other important factors: that is to say, the nature of the material in question and the aim of the research (as suggested by Armstrong, 1997: in Russo et al., 2012: 57). In particular, in the present study, the aim is that of analysing a database of intralingual and interlingual audio materials in an electronic format (subtitles generated in an asynchronous sequence), but also that of assessing if accessibility and accuracy requirements are met. Therefore, in order to avoid unnecessary complexities and to prevent transcription from being excessively time-consuming, it was decided to produce **basic reference transcriptions** to be used as gold standard for the subsequent analysis. This would not however prevent from adding annotations (*i.e.*, further information on background noise, interruptions, etc.) into the database or on the transcriptions in further studies having a different objective. This approach is also in line with automatic output generated by the reviewed ASR technologies, which, as it is described above, basically provide transcriptions with no punctuation at all (except for end-of-sentence full stops in *VoxSigma* and *Descript* app) and capitalization in proper names or at the beginning of the sentence (in *VoxSigma* and Descript app only). It is also worth mentioning the fact that hesitation or pauses are not included/indicated in the automatic transcriptions, nor in the gold standard material. For an overview of the annotation conventions adopted in the present study, Table 3.1 provides for a series of conventions largely based on EPIC transcription conventions.

Speech Feature	**Example from source**	**Transcription Convention**
Repetition	*Food food management*	food food management
Truncated words/hesitations	*Sin… Singapore;*	Sin… Singapore;
Empty pauses	Pauses or empty parts	Not transcribed
Abbreviations	*EP, FAO, UN*	EP, FAO, UN
Numbers	*3,000 tons; 2/3*	Three thousand tons; two thirds
Percentages	*30% of the population*	Thirty per cent of the population
Dates	*On 3 November of 2006; on November 3rd*	On 3 November of 2006; on November the 3rd
Unclear words/parts	When speech is unclear	(UNCLEAR)
Speech fillers	*"uhm", "em"*	"uhm", "em"
Speech markers	*Well, you know,* etc.	Well, you know, etc.
Exclamation mark	*!*	Not transcribed
Full-stop, question mark	. or *?*	Only at the end of a sentence

Table 3.1 - Transcription conventions adopted in this study.

Considering the very nature of the analysed material (*i.e.*, *impromptu* or read-out speeches), both for the purposes of speech-features representation and their subsequent analysis, this study strategy opted for reporting and transcribing all spoken expressions or words, both at a linguistic and disfluency level, including truncated words, mispronounced words, repetitions, etc. (see Table 3.1 above for a summary of speech features and elements included in the reference transcripts). The punctuation signs were specified only for end-of-sentence full stops and in case of question marks (when intonation is recognized by listening to speeches). In punctuation, however, some scholars argue that commas could play a very important role in the readability of texts. Commas could also disambiguate certain sentences, thus playing a very important role not only for readability, but also translation. The decision of not using them in the transcriptions (both in the ASR output and in the gold standard transcriptions) is based on the consideration that the ASR technology implemented (VoxSigma) do not make use of them unless explicitly added. As the present study intends to implement ASR technology under the standard, default configuration (by simulating a real time situation), commas were not incorporated.

Segmentation is based, as already said above, on the time-stamp segmentation generated by *VoxSigma*, without following the "unit of meaning" principle generally described in Interpreting Studies (Lederer, 1978): as already mentioned, the rationale of this decision is based on the assumption that, given the automatic nature of the workflow (like in a real time situation), the modification of the segmentation in units of meaning would be a sort of human intervention. In the case of GSR transcriptions, these were modified and aligned (manually with a work of *"copy&paste"*) so as to follow the segmentation structure reported by VoxSigma, allowing for a better comparison and analysis of both outputs. Disfluency elements such as speech fillers and speech markers were also included in the transcription.

As far as the spelling convention is concerned, the study's gold standard transcriptions mostly follow the standards applied in EU official documents, as indicated in the *Interinstitutional Style Guide* (European Union, 2020) available on the website of the Publication Office of the European Union. Additionally, it should be remarked that, provided that the aim of the present study is to assess whether accessibility requirements are met or not, an excessive weight of minor stylistic, conventional aspects of the language would undermine the final goal, that is to say to evaluate if ASR transcriptions are accurate. In this respect, it should be finally mentioned that the uttered abbreviations for proper names, institutions, organizations or official programmes/initiatives used internally or officially by the international organizations are transcribed *"as they are"* (approved conventional abbreviations). With regard to numerical values, all figures, values and percentages used in the source speeches were fully spelt out, except for dates that are expressed numerically (in line with the ASR output), as detailed in Table 3.1 above.

3.6. Taxonomy of Errors

After having described the database and the procedure followed in the configuration of the ASR system and in the processing of audio data, including the manual transcription of gold standard material, for next phase of the analysis, it was necessary to define an appropriate taxonomy of errors, which must be used to properly examine

ASR errors. This section of the chapter on methodology will thus offer a grid for the subsequent analysis of errors to be constructed on two different layers: **Coarse-Grained Errors** (**Layer 1)** and **Fine-Grained Errors (Layer 2**).

First of all, it should be pointed out that, for the construction of the grid, the taxonomy had to comply with two crucial requisites: *i.e.*, thoroughness and objectivity. As a matter of fact, if, on the one hand, it is necessary to identify the largest variety of error types (thoroughness), on the other, it is essential to adopt an objective approach in order to achieve conclusions and results which can be considered as sufficiently "objective" and "reliable" (these terms are between inverted commas as there may some criticalities in defining a 100% objective and reliable evaluation system).

Additionally, when defining the methodology for the errors taxonomy, it is indeed appropriate or recommendable to implement an objective strategy rather than exclusively relying on the subjective evaluation of a single evaluator. As already described in Chapter 2, during the last two decades several research projects were carried out on the application of ASR technology (namely, *Verbmobil*, *TC-Star* and *DARPA-GALE*, etc.) by mainly applying a quantitative method: see, for example, the works of Wahlster (2000) or Lazzari (2006). Also in the area of the studies for Accessibility purposes (see Chapter 2, §2.4), the statistical approach was mostly preferred to the qualitative methods. More specifically, in the scientific literature, to numerically quantify accuracy and thus the errors of ASR technology, the output assessment of intralingual live subtitling is generally based on the so-called *WER* (Word Error Rate) model, traditionally applied to the analysis of accuracy (see, for example, Dumouchel et al., 2011: in Romero-Fresco and Pöchhacker 2017: 150). Hence the statistical model-based accuracy becomes the measure or "meter" for the analysis and evaluation of the ASR output generated in the present study.

Upon these preliminary considerations, and the vast usage of this approach in similar studies, the adoption of a statistical, quantitative model is established here, together with the definition of a first layer of errors (Layer 1 – Coarse-Grained Errors) for measuring accuracy. In particular, the measurement of accuracy for ASR technology consists in the quantification of ASR technology "hits" or Perfect Matches (PMs), where "hits" or PMs represent the units of meaning or segments in a speech output which are perfectly matching with the reference (gold standard) transcription. According to McCowan et al., an ideal ASR evaluation system should in fact be:

(i) ***"Direct,*** *in other words, the measurement of the ASR component should be carried out independently of the ASR application,*

(ii) Objective*, the value of measure should be estimated or quantified in an automated manner,*

(iii) Interpretable*, that is to say the value of the measure should offer an idea about the performance, and, finally*

(iv) Modular*, the measure should be general to allow thorough application-dependent analysis*." (McCowan et al. (2005:2).

This study's coarse-grained taxonomy may prove to respond to these requirements and features. More specifically, to numerically quantify errors (representing the opposite value for "hits" segments), a taxonomic **Layer 1 (Coarse-Grained Errors)** was set up and identified for the purposes of offering a coarser taxonomy of errors. In particular, Layer 1 identifies three (3) main error typologies on a possibly objective way, by applying the scientific literature most used classification of errors based on the WER model. In this model, the first described error type of ASR technology is the complete omission or deletion of a word or more words in a speech (**Deletion**); secondly, the second type of error is the replacement of a word or more words with one or more different words (**Substitution**); and, finally, the third type of error is the addition of a word or more words which have not been uttered by the speaker in the source speech (**Insertion**). See Table 3.2 below for an example of each error type.

Error Type	**Description**	**Reference Transcription**	**ASR Transcription**
Substitution	Replacement of one or more words with one/more different words in the SR output	*"FAO has calculated that 20% of the population…"*	*"Foul has calculated that 20% of the population…"*
Deletion	Omission or elimination of one or more words from the source speech.	*"The emissions of CO2 have grown significantly in the last year"*	*The emissions of … have grown significantly in the last year*

Insertion	Addition of one or more words in the SR output.	*"The probability of controlling Climate Change…"*	*Of the probability of controlling Climate Change…*

Table 3.2 – Layer 1 for Taxonomy of Errors.

As already mentioned, the Word Error Rate (*WER*) model is the most popular measure for ASR evaluation in literature: it measures the percentage of incorrect words (Substitutions (S), Insertions (I), Deletions (D)) over the total number of words processed. More in detail, it is calculated according to the following formula:

$$\text{Accuracy rate } \frac{\text{N} - \text{Errors (D + S + I)}}{\text{N}} \times 100 = \%$$

Figure 3.9 – Formula for WER rate calculation

where N = total number of words, D = total number of deletions, S = total number of substitutions, I = total number of insertions.

Layer 1 (Coarse-Grained Errors) can therefore be considered as the main grid layer for the analysis and evaluation of accuracy in ASR technology output. It can respond both to the requisite of thoroughness and, possibly, to the requisite of objectivity (if backed by a remedy of validation described below: e.g. Inter-Annotator Agreement), making it possible to assess the ASR technology analysed here, but also investigating on the usage of other solutions in different contexts.

At this point, before describing this study's fine-grained taxonomy of errors, where an in-depth definition of errors is attempted and carried out, a series of considerations has to be made. First of all, it should be highlighted that the ASR system performance is dependent upon many different factors that could be grouped in the following categories by using the definition of Errattahi et al.:

- ***"Speaker Variabilities:*** *The ASR acoustic model may not be representative of all speakers in all their potential states. Variabilities may not all be covered, which affect negatively the performance of the ASR systems.*
- ***Spoken Language Variabilities:*** *The spontaneous and accented speech and the high degree of pronunciation variation are critical for ASR. Also, with large*

> *vocabulary, it becomes increasingly harder to find sufficient data to train the language models.*
>
> - ***Mismatch Factors:*** *The mismatch in recording or in technical conditions or in the media used as a source is the main challenge for speech recognition, especially when the speech signal is acquired on low quality conditions. The presence of background noise, the usage of poor-quality technology, the transmission channel and the recording devices can, indeed, introduce variabilities over the recording and decrease the accuracy of the system." (2018: 33).*

Apart from these considerations, it should be noted that quality assessment in intralingual live subtitling varies greatly across and even within countries. As explained by Romero-Fresco, what may be expected from these models of assessment is that they meet at least some of the following requirements:

> *"(1) They are functional and easy to apply,*
>
> *(2) they take into account not only the linguistic accuracy of the subtitles but also the comparison to the original speech,*
>
> *(3) they account for the possibility of reduced and yet accurate subtitles depending on the different national editing conventions,*
>
> *(4) they provide information about not only the accuracy of the subtitles but also other aspects of quality such as delay, position, speed, character identification, etc., (5) they account for the fact that not all errors have the same origin or impact on the viewers' comprehension; and*
>
> *(6) they provide an assessment of quality as well as an overall idea of aspects to be improved, in other words, food for thought as far as training is concerned". (2016: 57)*

Starting from these general considerations, a Layer 2 (Fine-Grained Errors) is defined for the taxonomy of errors. In particular, Layer 2 is based on a fine-grained classification of errors built upon five main categories: **Disfluency, Grammar, Lexis, Terminology,** and **Prosody**. For a detailed description of these categories and for some examples, including the reference literature, see Table 3.3 in the next page. The taxonomy so defined represents an attempt at condensing a wide range of speech features and error types identified in source speeches of this study.

FINE-GRAINED CATEGORIES	SPEECH FEATURE	EXAMPLE (FROM THE SOURCE)	REFERENCES IN LITERATURE
Disfluency	Acoustic variability	*coughing/applauses/laughing. background music, background noise, etc.* (ASR technology omitted these elements)	In Goldwater et al. (2010); Gada et al. (2013)
	Speech fillers	*ehm, uh, uhm, oh*, etc. (ASR technology omitted/added/replaced them)	In Ruiz et al. (2017); In Goldwater et al. (2010); In Adda-Decker and Lamel (2005); Gada et al. (2013)
	Speech markers	*"you know", "well", "so", etc.* (ASR technology omitted/added/replaced them)	In Goldwater et al. (2010)
	Hesitation/False start	*"Sin.. Singapore"; "we…when"* (ASR technology omitted/added/replaced them)	In Goldwater et al. (2010)
	Repetitions	*"Food food production"* (ASR technology omitted/replaced them)	In Goldwater et al. (2010)
	Start of speech/end of speech (omission or partial recognition)	*"Ladies and gentlemen"; "thank you"* (ASR technology omitted/added/replaced them)	In Goldwater et al. (2010)
Grammar	Tense form, grammar rules	*"define"* instead of *"defined"*	In Goldwater et al. (2010)
	Closed class word/Function words	pronouns, articles: *"their"* become *"these"*, etc.	In Goldwater et al. 2010; in Ruiz & Federico (2014)
	Negative form	*"can't"/"can"*	In Mirzaei et al. (2018)
	Contractions	*"it's"* vs. *"it is"*	Garofolo et al., 2004
Lexis	Lexical parts not recognized	*"afforestation"* becomes *"deforestation"*	In Fosler-Lussier & Morgan (1999); in Shinozaki and Furui (2001); in Gada et al. (2013)
	Multiple spelling of words	*"program" vs. "programme"*	Garofolo et al., 2004
	Numbers/dates	*"15%"* instead of *"50%"*	In Romero-Fresco & Pöchhacker (2017)
Terminology	OOV (Out of Vocabulary: Proper names, specialised terminology, abbreviations	*"FAO", "fall armyworm", GAFSP, etc.* (ASR technology omitted/replaced them)	In Romero-Fresco & Pöchhacker (2017); in Salimbajevs & Strigins (2015); in Gada et al. (2013)

Prosody	Intonation	*Question mark ("?"), Exclamation mark ("!")* (ASR technology omitted)	In Hirschberg et al. (2004); "prosodic variability"; in Goldwater et al. (2010)

Table 3.3 – Layer 2 for Taxonomy of Errors.

Before describing the set of rules used here to identify and classify the error types into five categories, it is fundamental to clarify that these categories are not intended to be objective or a complete classification of errors, as they may generate large margins of interpretation and not offer clear, unequivocal borders between two categories or among more categories. The high degree of ambiguity is for example evident in categories such as Lexis/Terminology or Grammar/Lexis. For example, the error *"afforestation"* (recognized as *"deforestation"*) may both be accounted for as a Lexis error and as a Terminology error. In this case, the correct option would be Lexis, as the world *"afforestation"* is not a specific term and it is part of the ASR system's vocabulary. Alternatively, with the substitution of the adjective *"their"* with *"them"*, the ambiguity between Grammar and Lexis does emerge. In fact, the misrecognition of the possessive adjective "their" can be examined both as the break of a grammar rule and as the replacement of a lexical element. Yet, as already done in most of the scientific literature produced so far, it may be useful to make some usage of these error descriptions (here referred to as Layer 2 Taxonomy) in order to better understand and assess ASR technology accuracy and, possibly, to correct the ASR system before implementing Neural Machine Translation, for further investigations.

Notwithstanding the large degree of subjectivity, a set of rules in defining the five categories in Table 4 above were established for the annotation of fine-grained errors. The categories below are made to generate a simpler, possibly less confused categorization of errors, though not eliminating ambiguity at all.

Now entering more into detail with regard to this classification, as far as the **Disfluency** features are concerned, it is possible to define that disfluency or speech disfluency includes speech-related or orality-related features like so-called *"false starts"*, *i.e.* words and sentences that are cut off mid-utterance; phrases that are restarted or repeated and repeated syllables; *"fillers"* or speech markers, *i.e.*, grunts or non-lexical utterances such as *"huh"*, *"uh"*, *"erm"*, *"um"*, *"well"*, *"so"*, *"like"*, *"you know"*, and *"hmm"*; and *"repaired"* utterances, *i.e.* instances of speakers correcting their own slips of the tongue or mispronunciations (before anyone else gets a chance to). Though speech disfluencies are widely believed to increase ASR error rates (for example in Goldwater et al., 2010; in Gada et al., 2013), there is little published evidence to support this belief. And this study may contribute to evaluate the weight of these errors.

With regard to **Grammar** features, the Fine-Grained Errors Taxonomy includes all errors related or connected with a wrong recognition by the ASR system for grammar rules or categories. An example of this is the error *"can't"* > *"can"* or *"him"* > *"he"*, where the rule for the negative form or the closed class of pronouns is misinterpreted by the system.

As far as the third category, **Lexis**, is concerned, it is just sufficient to clarify that, under this category, are all errors relating to lexical parts of the speech, thus including nouns and adjectives, as well as numbers and figures.

The present study takes an innovative approach because it accounts for specific terminology errors, under the **Terminology** category, an aspect which was not taken into account in previous works. As a matter of fact, specialized terminology is simply regarded as or classified as OOV errors, *i.e.*, Out of Vocabulary errors, without separating these errors from general categorizations of Deletion, Insertion or Substitution. Under this umbrella category, it is possible to find different errors connected with, or relating to names of institutions, international initiatives, domain-specific terminology and also proper names.

Finally, to complete the description of Fine-Grained Errors taxonomy, it is necessary to mention the prosodic features (**Prosody**) that are connected with speech prosody. In general terms, prosody is concerned with intonation, tone, stress and rhythm. For convenience, in this category, the study only takes into consideration errors connected with intonation, *i.e.*, with end-of-sentence question marks or exclamation marks.

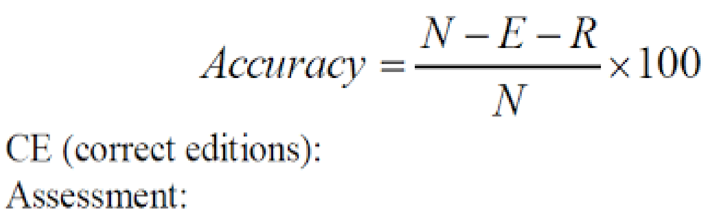
To further determine the methodology adopted in the present study, the model of statistical analysis selected for the measurement of ASR accuracy is described.

3.7. The NER model versus the WER model

In most of the scientific literature reviewed for the purposes of this study (see §2.4), the WER model described before represents the most popular instrument for the statistical quantification of errors in ASR system.

Although useful for assessing ASR output, this model is less precise in evaluating intralingual subtitling (Romero-Fresco and Pöchhacker 2017: 151), since it penalizes any error type with the same penalty, even when the meaning of the source file is retained. As already seen above, in the study, the three main Coarse-Grained Error typologies are Deletions (D), Substitutions (S) and Insertions (I). The measurement of accuracy would thus be the calculation of the number of total words, minus the number of errors (D+S+I), divided by the number of total words and then multiplied by 100. Under this system, each error would thus have the same *"weight"*. But, considering that a segment unit may continue to be fully understandable even if minor errors are present, for the purposes of accessibility and speech communications, a more-detailed evaluation of accuracy should be formulated, provided that it can anyway guarantee for a sufficient level of meaning and understanding in communications. The probable best response to this need is the so-called *NER* model.

Introduced for the first time in Romero-Fresco (2011) and developed further in Romero-Fresco and Martínez (2015), the model starts from the basic principles of the *WER* model, but it factors in the "seriousness" of errors and thus the effective subtitle quality (expression of accuracy measure). The *NER* model is one of a number of methods for determining the accuracy of live subtitles in television broadcasts and events that are produced using SR technology. The acronym three letters stand for *Number* (of words), *Edition* error and *Recognition* error. The model contains a formula to determine the quality of live subtitles: a *NER* value of 100 indicates, for example, that the content was subtitled entirely correctly. This overall score is calculated as follows: firstly, the number of edit and recognition errors is deducted from the total number of words in the live subtitles. This number is then divided by the total number of words in the live subtitles and finally multiplied by one hundred (see formula below in Figure 3.10).



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

CE (correct editions):
Assessment:

Figure 3.10 - Formula used by the NER model to calculate accuracy.

More specifically, *N* stands for the number words in the subtitles. Edition Errors (*EE*) are coincident with the "*result of the subtitler's strategic decision-making"* (Romero-

Fresco and Pöchhacker 2017: 152), but, in this study, Edition errors are not taken into account as our "subtitler" is a software solution and therefore there are no human decisions to evaluate. Finally, *R* errors are the recognition errors (D+I+A) which may be caused by mishearing and/or mispronunciation on the part of the ASR technology or by other factors. Again, these errors may be deletions, insertions or substitutions (as in the WER model).

If comparing the *NER* model with the *WER* mode, it should be highlighted that the latter is static, since it simply measures the textual discrepancy between that which was written and spoken without evaluating the seriousness of errors. The most important element of innovation in *NER* model is the fact that the macro error types (Deletion, Insertion and Substitution) are classified as minor, standard or serious errors. But, for convenience, in this study, we have "weighted" the reported errors only as "Serious" or "Not Serious", giving a score of 0.5 to "Not Serious" and 1 to "Serious" ones, respectively. For the statistical evaluation of accuracy, Coarse-Grained Errors taxonomy only is taken in consideration. Despite the use of new labels, the error-grading system partially remains the same as in the NER model. More in detail, *"Not Serious"* errors cause a certain loss of meaning, without compromising the meaning and content or the understanding of the segment or subtitle unit. On the contrary, *"Serious"* errors deprive the viewer of a correct understanding of an idea unit, the source-text content being lost, including a change of meaning of the source text. This type of errors also introduces "*factual mistakes or misleading information"* (Romero-Fresco and Pöchhacker 2017: 152) that could make sense in the new context. A certain degree of subjectivity is certainly associated to the process, but, as defined in the next section of this chapter and in Chapter 4, the taxonomic scheme and error grading system implemented here will be validated by means of inter-annotator agreement. For examples of *Serious* or *Not Serious* errors, see Table 3.4 below.

SR Output	**Reference Transcription**	**Error-grading**
"The government has reduced public spending by 15%"	*"The government has reduced public spending by 50%"*	Serious, weight score: 1
"(Deletion) FAO has expanded investments in Africa"	*"Well, FAO has expanded investments in Africa"*	Not Serious, weight score: 0.5

Table 3.4 – Error grading system.

In the calculations of the WER and NER rates, this study implemented a method of calculation partially adapted to the fully automatic features of the ASR system deployed, where human intervention is not included (except for the evaluation process of annotation data commented above): in fact, the role/contribution of a respeaker is not considered here. Additionally, it should be clarified that the most relevant rate for the present study is the NER rate, as it accounts for the *"Not Serious/Serious"* error severity classification described above. Furthermore, under this study, the NER rate was broken down into two different NER rates, which are renamed NER1 and NER2 so as to include or exclude "Not Serious" errors from the calculation, respectively. Therefore, the accuracy NER1 rate will include the occurrences of Not Serious errors, while NER2 rate will exclude those errors totally. This should help in better representing the severity differentiation of errors and in responding more efficaciously to the various applications of live subtitling (inter-lingual and intralingual subtitling for non-hearing people and NMT application). In fact, with the NER2 rate it is possible to examine accuracy without counting for the minor errors of a given segment. Minor errors do not alter the overall meaning (and the understanding) of the segment unit. In the NER1 rate, *"Not Serious"* errors are assigned with a penalty of 0.5 points (*"Serious"* errors have a 1 point penalty), while in NER2 rate, *"Not Serious"* errors are not considered at all in the formula used for the calculation of accuracy. NER2 adapted model includes in the weighing system the possibility to not penalise an 'error' based on the fact that it does not worsen the output. For example, the omission of disfluencies (*"uhm", "um", "well"*) may not alter the readability of the subtitles, and it is something that many ASR systems implement to 'clean up' the transcript as much as possible. In the default configuration of the software solutions implemented here, the disfluency elements are given. Implementing an evaluation model (like NER2) where it is possible not to penalise a minor element that is transcribed by ASR (i.e. by giving such shift a zero weighing) may help solve this potential issue.

At the conclusion of this section, it is important to underline that *Serious* errors could hamper the understanding of the final output and of the entire speech. As we are also considering the possibility of applying Automatic Machine Translation (based on NMT engine) in the subsequent phase of the analysis, these types of errors would cause deviations and serious errors in the target interlingual subtitles.

Generally, in the subtitles industry and in literature, accuracy for subtitles is measured and considered as acceptable when subtitles achieve a score of at least 98% with the WER (a few countries and TV broadcasting authorities continues to use this rate) or the NER model (the largely-used model). However, further considerations should be made with respect to the potential application and usage of ASR and NMT technologies together, as better discussed in the conclusions of Chapter 4. Before considering the evaluation of accuracy in this study's final transcriptions and NMT output, it is necessary to validate, within the methodological framework, the taxonomic scheme previously defined. To do so, a validation method should be achieved through what is called in literature as an inter-annotator agreement test, as described in the section below.

3.8. Inter-Annotator Agreement Test

In computational linguistics and, in particular, in speech corpora analysis, the usage of annotations represents an important tool to analyse audio/video material and to make specific comments or add detailed information on a set of texts (see, for example, Bendazzoli, 2010: 76). Yet, before continuing with the categorization and analysis of the project data, a series of considerations should be done. First of all, it should be stated that:

> *"The building up of linguistic resources, and, more generally, the annotation of data, imply the formulation of subjective judgements or evaluations. The necessity of establishing the extent to which these evaluations can be reliable and reproducible has gained increasing importance, and has made the validation process a consolidated practice". (Gagliardi, 2018: 1; my translation)*

The taxonomy defined in §3.6 above should therefore be evaluated so as to assess whether it is reproducible by other annotators or evaluators – and hence sufficiently reliable. Hopefully, this will make the annotation process adopted in this research a consolidated and largely accepted practice by other researchers. This is particularly important for the Coarse-Grained Error categories of Layer 1 (Deletion, Substitution and Insertion) and for the pair Serious/Not Serious errors, as these parameters have effects on the calculations made in relation to the accuracy of software transcriptions.

Another important consideration to be made regards the very nature of the annotation system adopted here. Given the typology and complexity of the audio and video contents that do not allow for the usage of an automatic annotation system, this study is mainly based on manual annotation. But, if on the one hand, manual annotation *"allows exhaustive and detailed corpus-based analyses [...] that would not be possible with purely automatic techniques"* (as indicated in Fuoli and Hommerberg, 2015: 316), on the other, it should be remarked that the taxonomic validation may be a complex and, above all, a subjective task. And again, by using the words of Fuoli and Hommerberg (2015: 316): *"this may hinder the transparency, reliability and replicability of analyses"*.

More in particular, the study implemented an approach to taxonomic validation based on two specific strategies. Firstly, a series of annotation instructions was defined and drafted in a sort of Annotator's Manual to be made available to other annotators (7 annotators, plus the author of this study). Secondly, the reliability and replicability of the annotation procedure was validated by using a special instrument, the so-called Inter-Annotator Agreement Test. At the basis of this test is the **Inter-Annotator Agreement** (abbreviated as IAA), which is described as follows by Gagliardi:

> *"Within the computational context, IAA is used as a means to pass from annotated material to a gold standard that is a set of data which is sufficiently noise-free to be used for training and testing purposes". (Gagliardi, 2018: 1; my translation)*

In general terms, and for the purposes of the present study, the inter-annotator agreement is mainly a measure of the extent to which the annotators (selected in first phase of the test) make the same decisions when assigning pre-defined categories to the different segment units of the text (on the basis of the project's taxonomy defined above in §3.6). Also Artstein and Poesio (2008: 557) confirm this: *"data are reliable if coders can be shown to agree on the categories assigned to units to an extent determined by the purposes of the study"*.

As mentioned above, the first phase of this test process was the drafting of a series of instructions (the Annotator's Instructions available in **Appendix B**) to be provided to a number of voluntary participating annotators. This sort of annotator manual was created by meeting two main criteria: simplicity and clarity. Given that the annotators involved in the test are not specialized researchers in the field of ASR, nor experts in audio/visual annotation or in ASR transcription, the instructions were defined by keeping in mind that the target readers had no specific skills in that field. The choice of not involving ASR experts or ASR scholars was based on the idea that IAA results could be considered even more robust if high rate were reached in the end of the test. For this reason, the instructions were to be formulated in a simple and not-ambiguous manner, offering maximum clarity in terms of taxonomy, definitions and concepts. Practical examples were also included for each set of instructions. In order to be as simple as possible, the manual had also to be short in terms of number of words, and use no specialized terminology from the computational linguistics sector for a better readability. The Annotator's Instructions were also accompanied with a descriptive e-mail regarding the test. The full version of the Annotator's Instructions can be consulted in the section "Annotator Kit" of **Appendix B**.

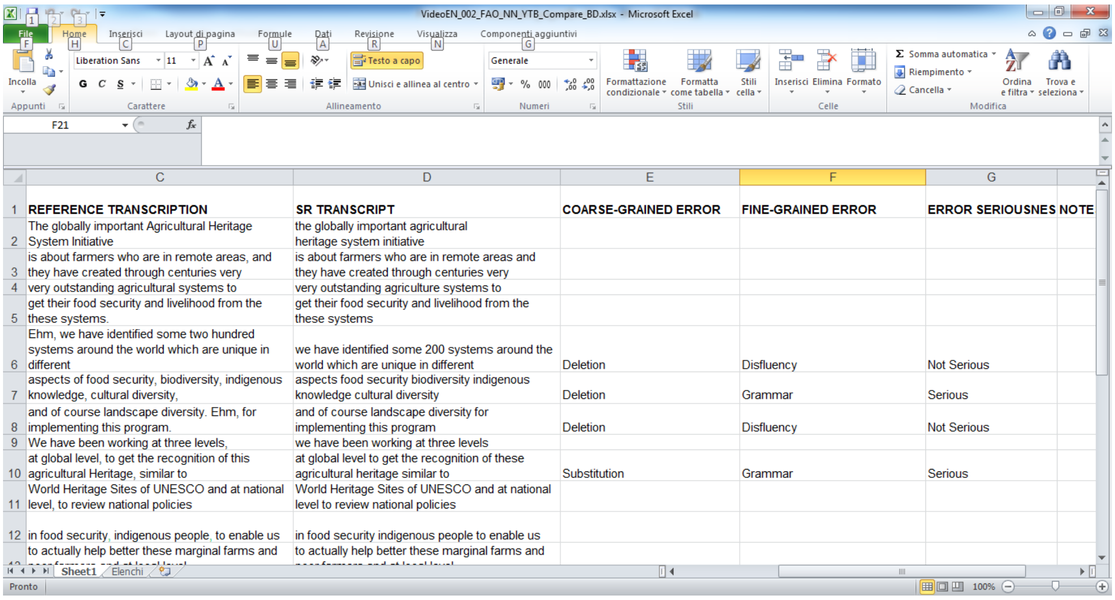
The second phase in the process was characterized by the decision as to how many annotators to involve in the annotation task, and how. According to Spooren and Degand (2010):

> *"There are three main strategies that can be applied in situations where reaching high inter-coder agreement between independent coders is challenging:*
>
> - *Double coding*
> - *Partial overlap between two or more coders.*
> - *One coder does all." (in Fuoli & Hommerberg, 2015: 17-18).*

In the *Double coding* method, two annotators annotate the entire set of data independently and then discuss all the disagreements until full consensus is reached. In the second method, a portion of the data is annotated separately by two annotators, while the rest of the corpus is annotated by only one person. The Agreement rate in this case is calculated on the limited portion only. Finally, in the *One coder does all* method, the entire corpus is annotated by only one annotator. This may be the most subjective modality as the annotator can apply the predefined categories in an autonomous way, so the reliability of the annotation scheme may prove weak. Given these considerations and requisites of objectivity and reliability, and also the very nature of this study, it was decided to implement the second modality above for the inter-annotator agreement calculation, where the annotators work on a restricted set of texts and where the inter-annotator agreement is calculated on the basis of the data provided by the main annotator of the project and the other 7 annotators, in relation to a restricted sample of texts only.

The external annotators involved in the testing phase included researchers/PhD students working and studying in the linguistic field, all coming from the Department of Interpreting and Translation (University of Bologna). The participants were not directly engaged in the present research and included 6 female individuals and 2 male individuals with an age ranging from 25 to 50 years old (with 7 annotators of Italian nationality and 1 of Chinese nationality; all annotators had English as L1 or L2). All annotators were translators/interpreters and/or linguistic experts.

After these preliminary steps, the testing phase started with the distribution of the Annotator Kit via E-mail including the Annotator's Instructions manual, and the two sample files extracted from the database (file 002 and file 012, for a total duration of about 4,3 minutes of speech) to be completed and annotated, after reading the instructions. The two sample files were chosen according to the criteria of database representativeness. In particular, they include both a Native-speaker file and a Non-Native-speaker file (generated by VoxSigma and Google Speech Recognition via YouTube). The two files to be annotated were provided in the form of an Excel file where the reference transcription and the software transcription were already supplied for (see an example below in Figure 3.11). The rationale for this stands in the possibility for annotators to annotate the transcriptions more rapidly and to work on the same material.



	C	D	E	F	G	
1	REFERENCE TRANSCRIPTION	SR TRANSCRIPT	COARSE-GRAINED ERROR	FINE-GRAINED ERROR	ERROR SERIOUSNES	NOTE
2	The globally important Agricultural Heritage System Initiative	the globally important agricultural heritage system initiative				
3	is about farmers who are in remote areas, and they have created through centuries very	is about farmers who are in remote areas and they have created through centuries very				
4	very outstanding agricultural systems to	very outstanding agriculture systems to				
5	get their food security and livelihood from the these systems.	get their food security and livelihood from the these systems				
6	Ehm, we have identified some two hundred systems around the world which are unique in different	we have identified some 200 systems around the world which are unique in different	Deletion	Disfluency	Not Serious	
7	aspects of food security, biodiversity, indigenous knowledge, cultural diversity,	aspects food security biodiversity indigenous knowledge cultural diversity	Deletion	Grammar	Serious	
8	and of course landscape diversity. Ehm, for implementing this program.	and of course landscape diversity for implementing this program	Deletion	Disfluency	Not Serious	
9	We have been working at three levels,	we have been working at three levels				
10	at global level, to get the recognition of this agricultural Heritage, similar to	at global level to get the recognition of these agricultural heritage similar to	Substitution	Grammar	Serious	
11	World Heritage Sites of UNESCO and at national level, to review national policies	World Heritage Sites of UNESCO and at national level to review national policies				
12	in food security, indigenous people, to enable us	in food security indigenous people to enable us				
	to actually help better these marginal farms and	to actually help better these marginal farms and				

Figure 3.11 – Example of one Excel file provided to annotators.

The predefined categories for column E (Fine-Grained Error), F (Coarse-Grained Category) and G (Error Seriousness) could be easily entered by selecting the correct option from a drop-down menu including the set of predefined categories only. At the end of their annotation process, annotators then sent their outputs (the two Excel files) to the evaluation phase for the estimate of the inter-annotator agreement rate. As described below in Chapter 4 on the Analysis of data, the calculation of the agreement rate will be broken down at different levels in order to evaluate the inter-annotator agreement rate achieved, more in detail.

To conclude the description of the methodological framework, and to better understand what is quality and accuracy in the workflow ASR+NMT examined here, the presentation of the statistical model used to evaluate the accuracy rate in Neural Machine Translation output (the third step in this study pipeline, see Figure 8 above) is offered in the following section.

3.9. Application of NMT: the NTR model

When defining the methodology for the final phase of the ASR+NMT pipeline, it was decided to apply the Neural Machine Translation (NMT) only to a limited number of files having achieved a high accuracy level in ASR evaluation. In other words, NMT will be applied only to those files having met the minimum accuracy requisite of 98%. The NMT solution selected for this part of the study is *DeepL*[29], a popular, marketed Neural Machine Translation solution capable of meeting the requisites defined in literature (see Chapter 2, §2.3.3) for Neural Machine solutions. However, after obtaining the target subtitles in the target language of the present study (Italian), the necessity of measuring accuracy of these subtitles emerges. To do so, a statistical model is defined here. The model is an adapted version from the statistical method used in Romero-Fresco and Pöchhacker (2017: 159) and it is denominated *NTR* model, where *N* stands for Number of words, *T* for Translation errors and *R* for Recognition errors. The accuracy is calculated by using the formula shown in Figure 3.12 below.

[29] DeepL: https://www.deepl.com/translator

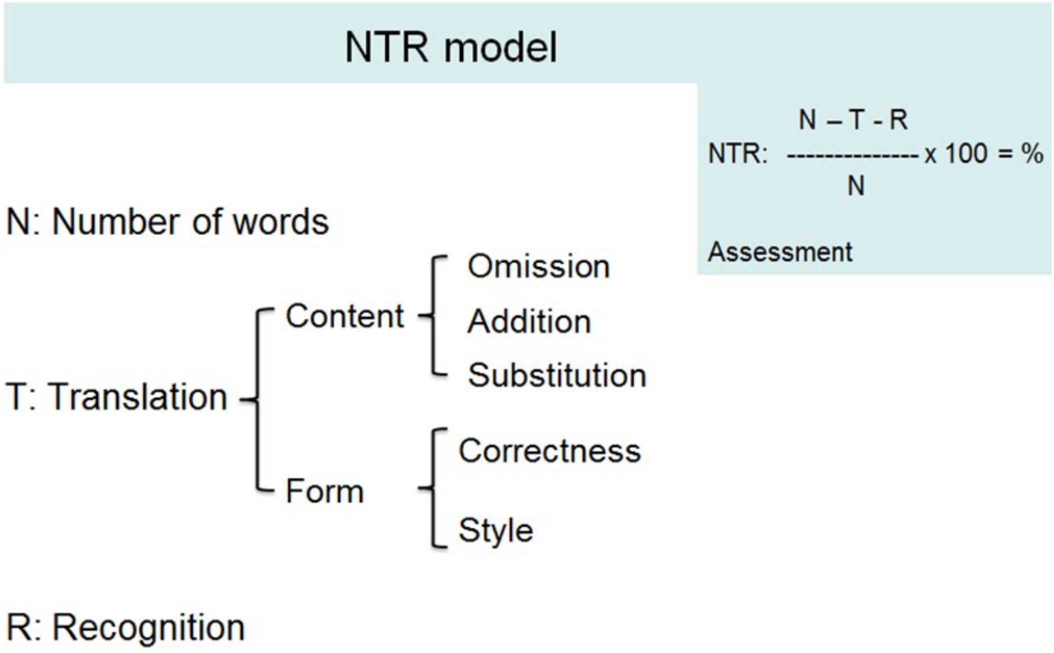


Figure 3.12 – NTR model definition and formula (Romero-Fresco and Pöchhacker, 2017).

More in detail, in the methodology for the NMT accuracy evaluation, it was decided to adapt the model above so as to establish a classification of errors based on three different layers of analysis **Coarse-Grained Errors**, **Fine-Grained Errors**, and **Error Severity**. For the first layer, the present study maintained the category classification adopted in §3.6 for the Automatic Speech Recognition, *i.e.*, Deletion, Substitution and Insertion. During the annotation process, it is therefore indicated a Deletion error when the NMT technology omits a term or series of terms in a given segment unit; a Substitution error when the NMT replaces a term translation with a wrong term or a series of terms translation; and, finally, an Insertion error when the NMT adds one or more terms which are not present in the reference source speech.

Unlike the previous taxonomic scheme implemented for the Automatic Speech Recognition assessment, for the Fine-Grained Error categories this study adopted a distinction based on two categories only: namely, Content and Form (Romero-Fresco and Pöchhacker (2017: 159). *Content* errors can be omissions, additions or substitutions (typically mistranslations by the NMT software) relating to the loss of information (for example, wrong numbers or a missing term bearing a significant piece of information); *Form* errors can affect the correctness of the subtitles in terms of grammar or their style (appropriateness, naturalness, register). All errors are then classified by their degree of severity using a three-level grading system and scoring system in line with the approved terminology of the LISA QA metric[30] (property of the Localization Industry Standards Association): according to this categorization, error seriousness is classified as "Major" when the error has a major impact on the subtitle unit, "Minor" when the error has a low impact on the understanding and accuracy of the subtitle unit and, finally "Critical" when the error seriously compromises the understanding and meaning of the segment unit. Examples of these error categories will be provided later on in Chapter 4.

As far as the methodology is concerned, a few considerations should be now offered here to further understand the analysis of NMT output carried out in Chapter 4, especially in consideration of the implications derived from the combination of ASR with NMT. As already highlighted in previous studies (see, for example, Ruiz and Federico, 2014, or Goldwater et al., 2010), an "*increase in WER rate in ASR can significantly increase the so-called Translation Error Rate (TER) in the NMT output"* (Ruiz and Federico, 2014: 4). Again, as suggested in Ruiz and Federico (2014), the analysis suggests that *"substitutions have a greater impact (on translation quality) than deletions or insertions"* (ibid). In particular, it is interesting to observe, together with Goldwater et al. (2010), that different implications are generated in the ASR-NMT pipeline when the AST system encounters what Goldwater et al. (2010: 182) calls *"function words"* (also known as closed class words) and content words. The former group of words is much *"more problematic for speech recognition"* (Ruiz and Federico, 2014: 10). As a matter of fact, using the words by Ruiz and Federico:

> *"The speaker may alter the pronunciation of high frequency function words, such as prepositions and articles, by under-articulating or dropping phonemes. While a human can predict these words with high accuracy, an ASR system relies on phoneme or triphone recognition as an intermediate step toward recognizing words". (Ruiz & Federico, 2014:10)*

However, in the present study's NMT output, it will be necessary to verify if this kind of ASR errors generates errors with a Minor or higher grading in NMT transcriptions. Another group of words, which Goldwater et al. (2010: 198) define as Content words

[30] LISA and its marks are the property of the Localization Industry Standards Association and are used with permission.

(also known as open class words), can be described in their role within the ASR+AST pipeline as follows:

> *"…are generally simpler to recognize, as they often contain more syllables and cover a larger amount of speaking time within an utterance. On the other hand, open class words might not be represented in a speech lexicon, rendering them impossible to be generated by an ASR system". (Ruiz and Federico, 2014: 11)*

Yet this group of words may also prove to be more problematic in this study's evaluation. In fact, as demonstrated by Vilar et al. (2006) in a study on ASR and SMT (Statistical Machine Translation), *"missing content words contribute more toward translation errors than missing function words"* (Ruiz and Federico, 2014: 10). And similar considerations can also be applied to this study though it focuses on NMT (and not on SMT). It is also interesting to observe that Terminology errors (or OOV – Out of Vocabulary as defined in §3.6 above) in ASR may often be the cause of Content errors in NMT, with a Critical error grading, especially in the case of Substitution and Deletion errors in ASR, as also reasoned in Ruiz and Federico:

> *"Substitution errors on content words, however, have a significantly lower impact. Conversely, deletion errors on content words have a greater impact than those on function words." (Ruiz and Federico, 2014: 11)*

The analysis carried out in Chapter 4 will also examine the "weight" of terminology-related errors on the final output in Italian language.

3.10. Summing up

To conclude this chapter on methodology, it is possible to underline that the present study is based on a specific scope of research (speeches on climate change) and it starts from research questions that may contribute to expand the horizons of research in the field of ASR technology and NMT. The principles of representativeness and authenticity have been the key in the definition of the database and in the organization of the information and data collected here. A detailed description of the workflow for the database building phase and the audio/video material processing was supplied for, together with a series of requisites that were met in selecting the appropriate ASR technology. In this respect, it was possible to see what are VoxSigma's and Google

Speech Recognition technology's features and their function in the methodology implemented. After that, the criteria and conventions followed for the compiling of reference transcriptions were defined for the purposes of establishing a possible protocol for any potential transcriber or annotator in a scenario similar to that analysed in this study.

After setting up a possible ASR+NMT pipeline and defining the protocol for data processing, the taxonomy of errors was determined and created, organizing it into two layers (Layer 1 - Coarse-Grained Errors, and Layer 2 Fine-Grained Errors). Trying to meet the criterion of objectivity, a simpler, general Layer 1 for taxonomy was established, including three categories of errors only (Deletion, Substitution and Insertion). Secondly, in order to offer a sufficient representation of ASR errors typologies, further categories of errors were identified and described more in detail: Lexis, Grammar, Terminology, Disfluency and Prosody. Specific considerations were also made with respect to the possible different interpretation and the ambiguity associated to, or connected with these five categories.

Finally, a comparison between the WER and NER models was carried in order to design a better statistical approach, and a modification of the NER model was proposed to better adapt it to this study's scenario. All these considerations and suggestions will then be tested and experimented in the following phase of the study (see Chapter 4).

Furthermore, thanks to the data processing protocol elaborated in the present study, it is possible to set the basis for assessing the performance of ASR software applications, allowing public institutions or international organizations (like the ones indicated in the study scenario) to effectively identify the software solutions which better respond to their needs. Some public institutions could, for example, prefer a ASR solution capable of "recognizing" specialized terminology (with the possibility of training the software to it), while others may choose ASR software that can better cope with issues related to the Non-Native variable (a less specialized terminology, but with a higher level of prosodic and regional-specific speech elements). The use of a statistic model can also help public institutions in determining whether the usage of AI-based solutions can (or cannot) partially replace human resources when professionals are not available for certain language combinations. The most important aspect of the methodology defined in this Chapter probably concerns the possibility of evaluating accuracy in subtitles produced for conferences or speeches on climate change held in real-time (with subtitles created in an asynchronous way).

4. Data Analysis and Discussion of results

4.1. Introduction

In this chapter, a detailed analysis of transcription data collected and generated through the Automatic Speech Recognition (ASR) process and the Neural Machine Translation (NMT) is conducted, covering the entire pipeline described in Chapter 3. In particular, the analysis will be structured and formulated according to the different components and elements composing the present study's data and methodology. Firstly, a quantitative description of audio/video transcription material will be provided in order to better understand the composition and nature of the audio/video files submitted to ASR (see §4.2 below). Secondly, the results of the Inter-Annotator Agreement Test conducted in the earlier experimental phase will be discussed and presented in order to validate the taxonomic scheme used to annotate the transcriptions generated by ASR applications (see §4.3). Thirdly, an in-depth analysis of *VoxSigma*'s transcriptions will be presented to better grasp the composition and errors distribution of ASR (§4.4). In particular, for this part of the study, a detailed analysis of data per different taxonomic categories will be presented, both for Native and Non-Native speaker files, trying to underline criticalities in automatic speech recognition. More specifically, the study will determine the statistical percentage values for the Deletion, Substitution, and Insertion errors (§4.4.1) in relation to the Native/Non-Native variable. The same approach will be applied to the categories of errors belonging to the present study's taxonomic Layer 2: *i.e.*, Lexis, Grammar, Terminology, Prosody and Disfluency (see §4.4.2 below). Though less important to the measurement of accuracy and the calculation of the WER and NER indexes, these categories can however prove to be useful for the purposes of describing the software behaviours and the ASR system's potential criticalities. Fourthly, the analysis will focus on the categories of "Serious/Not Serious" errors (see §4.4.3) which, as already specified in Chapter 3, have a major function in the statistical model used (WER, NER) and, accordingly, a major impact on the measurement of accuracy. The weight of these errors on the final output of the software's automatically generated transcriptions will be discussed and examined. At the end of the analysis, an evaluation of accuracy will be attempted (see §4.5) according to the definition of accuracy given in Chapters 2-3, by examining the WER and NER rates, both for Native and Non-Native speaker files. After this step, a comparison of the outputs from both software solutions (*VoxSigma* and *Google Speech Recognition*) for a limited sample of files (§4.6) will be carried out. Subsequently, the study will analyse (in §4.7) the automatic speech transcriptions automatically translated by the previously-selected solution of Neural Machine Translation (namely, *DeepL*). This analysis will be carried out for a limited number of Native speaker transcripts having recorded a high-accuracy level according to the previous analysis step. At the end of the analysis, the study will examine the application of Augmented Terminology (AT) resources (§4.8), that is to say specific terminological resources used to enhance the ASR output in an unprecedented strategy with respect to the reference literature. This will be carried out, firstly, by collecting specific approved terminological resources created by the Food and Agriculture Organization (FAO) and, secondly, by applying those resources to the automatic recognition process. The impact of Augmented Terminology (AT) will then be calculated in terms of accuracy for two audio files where domain-related terminology has a significant weight on the percentage of errors (as per Taxonomy - Layer 2). The chapter concludes by summarising the main results through a Discussion of results (§4.9), while trying to define possible improvements or strategies that may help in tackling the difficulties and problems encountered in the present ASR+NMT system experimentation.

4.2. Quantitative description of data

For a quantitative description of the present study's database (the complete database is available in the Appendix A for consultation), after having partially described it in Chapter 3 (§3.3) on methodology, it should be recalled that it includes 55 audio/video files containing official speeches on climate change and its effects on agricultural production. All speeches were held by international experts, officials or politicians (in a mono-speaker format or as read-out presentations) at international conferences or institutional debates which were hosted by non-governmental and governmental organizations between 2013 and 2019. The corpus of audio/video texts amounts to a total of 44,838 words[31] and a total of 5 hours, 53 minutes and 34 seconds[32]. The

[31] The total number of words is calculated on the basis of the total words number of the Reference Transcription material for this study and it was obtained by using Microsoft Excel spreadsheet calculation.

[32] The total duration of the audio/video material is calculated on the duration of source files, excluding any cut portions.

average length of each video/audio file is of 6 minutes and 26 seconds (with a minimum duration of 1’ 11’’ for file 041 and a maximum duration of 25’ 23’’ for file 048). The speakers are from 34 countries and they can be divided into two different groups: Native and Non-Native speakers, as already defined in Chapter 3 (§3.3.2). In **Appendix A**, it is possible to have an overview of the speaker database composition and their country of origins, together with an indication of their gender. In particular, it is possible to see that the database includes 50 speakers from 34 countries. If the texts distribution is broken down by country of origin of the speaker, it is possible to see that the database includes 6 speeches from the United States; 4 speeches from the United Kingdom and Brazil, each; 3 speeches from Spain and Ireland, each; 2 speeches from Australia, Sri Lanka, Bangladesh, Ghana, Ireland, Belgium, Romania, and the Netherlands, each; and 1 speech from all of the remaining countries (Portugal, Iceland, Namibia, Hong Kong, Liberia, Sweden, Pakistan, South Korea, India, Lesotho, Republic of Nauru, Slovakia, South Africa, Swaziland, Iran, Jamaica, Ethiopia, Poland, Indonesia, Germany and Norway), each. If the speech distribution is analysed further, it is possible to observe that the gender composition is as follows: 45 speeches are held by Male speakers, while 10 speeches are held by Female speakers. At this point, if the database is subdivided according to the Native/Non-Native categorization, it is possible to see that the distribution of the speaker population per minutes of speech is as follows in Table 4.1 and in Figure 4.1 below.

Group	**hh:mm:ss**	**Percentage**
Native	02:49:43	48%
Non Native	03:03:51	52%

Table 4.1: Native/Non-Native composition of the speaker population per minutes.

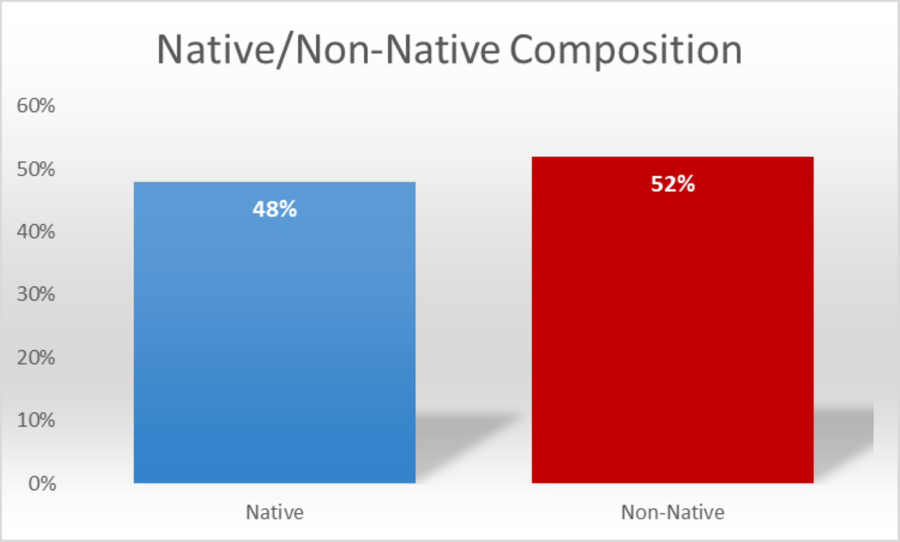


Figure 4.1 – Native/Non-Native composition of the speaker population per minutes.

An approximate similar distribution of the speaker population can be found if the number of total words as per the groups of Native and Non-Native speakers is examined: 25,074 words from the Non-Native group, and 19,764 words from the Native group, respectively.

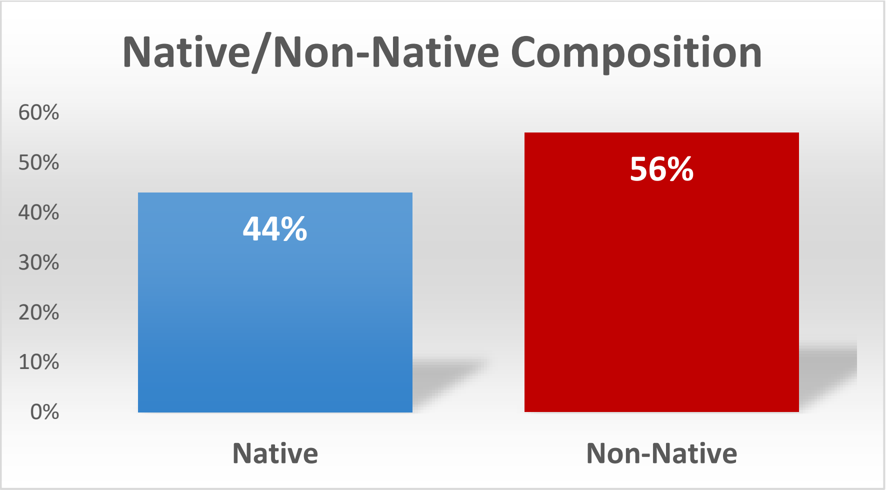


Figure 4.2 – Distribution of the speaker population per Native/Non-Native based on the number of words.

In Figure 4.2 above, it is possible to observe the distribution of the transcription database based on the number of words. In this case, the percentage of Native speaker words is 44% (lower than the value that was obtained in the per-minutes distribution), while the percentage of Non-Native speaker words is 56%. Now, if the Gender categorization is examined, it is possible to highlight that the male speaker variable is predominant across this study's population, and this is mainly due, among other reasons, to the fact that politicians and officials representation at international organizations generally sees a prevalence of male individuals (see, for example, the report by ISPI, 2012). To show the remarkable disparity in gender composition for the present study's database, in Figure 4.3 below, the percentage of Male/Female speakers is indicated based on the per-minutes representation of the population.

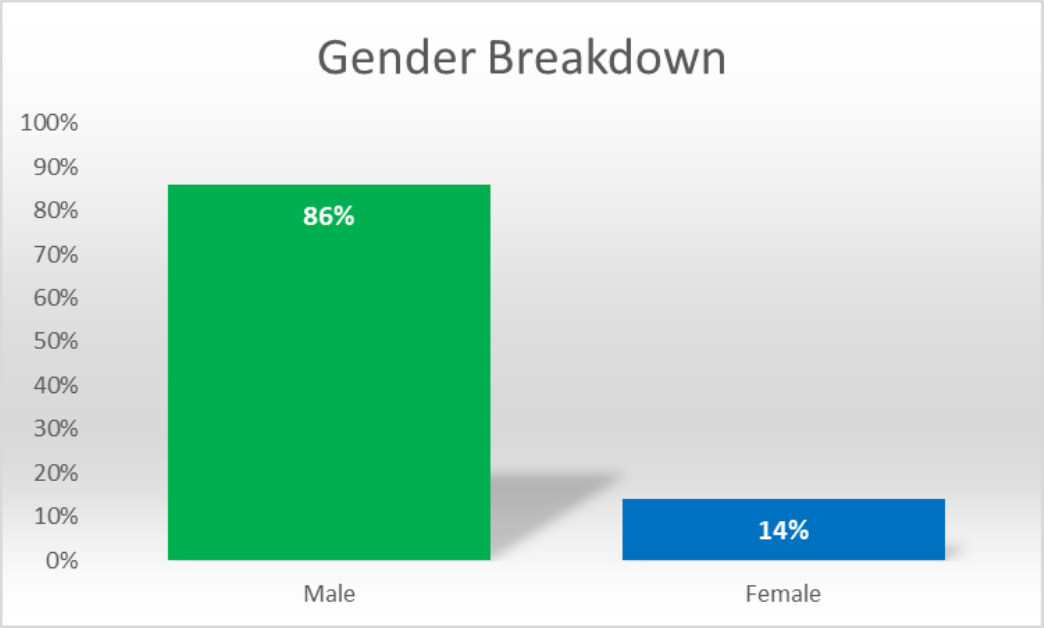


Figure 4.3 – Male/Female representation in the database per minutes.

Finally, it is possible to describe the database by observing the speaker population according to the speech speed variable. For this variable, the database includes 22% of the speech minutes at a Slow speed rate (as seen in Chapter 3, a slow speed rate is a speed value below 110 words per minute), a 58% of the sample with an Average speed rate (between 110 and 150 words per minute) and, finally, 20% of the speech sample with a Fast speed rate (over 150 wpm), as shown in Figure 4.4 below.

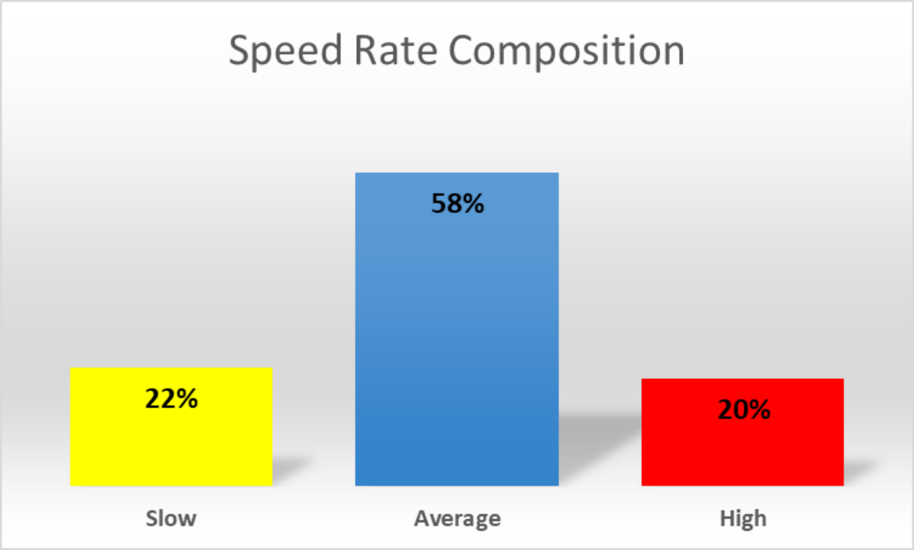


Figure 4.4 – Composition of the speaker population according to the speed rate (wpm).

For this variable too, it is possible to claim that the database population is sufficiently balanced for the purposes of this study, where a little more of the half of the sample has an Average speed rate, and the other half of the population has a Slow or Fast speed rate.

Finally, when examining the speed rate by Native/Non-Native distribution it is possible to see that the Native speakers have a higher average speed rate if compared to Non-Native speakers, as shown in Table 4.2 below.

Group	**Speed Rate (words/m)**
Native	138.73
Non Native	125.11

Table 4.2 – Average speed rate by Native/Non-Native group of speakers.

4.3. Results of the Inter-Annotator Agreement Test

Before quantitatively examining the errors distribution and the accuracy of ASR transcriptions, it is necessary to validate the taxonomic scheme described in Chapter 3 on Methodology (§3.6). As already mentioned, this validation is made possible by carrying out an Inter-Annotator Agreement Test (as defined in §3.8), the results of which are analysed here. More specifically, the calculation of the agreement rate was broken down at different levels in order to evaluate the inter-annotator agreement rate in detail. As already seen in §3.8, 8 annotators took part in the test (the main annotator and further 7 extra-project annotators). They represent a mixed pool of annotators who work and belong to different scientific, sub-disciplinary areas of studying in the area of Linguistics and Translation/Interpretation.

First of all, the test results were examined to calculate the agreement rate in relation to the presence/absence of errors for each segment unit in both sample files (the complete procedure is described in §3.8). To do so, the agreement rate was calculated by comparing the annotations in relation to the Perfect Matches (PM) in the files. In the study, a segment unit is considered to be a Perfect Match (PM) when the text of the reference transcription is identical to the transcription generated by the software, without considering the differences in punctuation and the differences in uppercase/lowercase letters. For file 002 of the test, the average inter-annotator agreement rate about the presence/absence of errors (Perfect Match) so obtained was 89% (here rounded-up for convenience), as shown in Table 4.4 below. Starting from the left, the column "Perfect Match" (reference) includes an indication of whether the segment generated by ASR is a perfectly identical to the reference transcription (gold standard). The column *"# of annotators spotting mistake in segment"* ("Yes/No") reports the number of annotators recognizing the presence/absence of error with respect to the Perfect Match. Finally, the column *"Agreement with reference (%)"* contains the rate of agreement among annotators is specified as a percentage value (%). At the bottom of Table 4.4, the average IAA rate is reported (rounded-up average value).

Segment Unit	**Perfect Match (reference)**	# of annotators spotting mistake in segment		**Agreement with reference (%)**
		Yes	**No**	
1	**Yes**	8	0	**100%**
2	**Yes**	8	0	**100%**
3	**No**	4	4	**50%**
4	**Yes**	8	0	**100%**

5	**No**	0	8	**100%**
6	**No**	4	4	**50%**
7	**No**	1	7	**87.50%**
8	**Yes**	8	0	**100%**
9	**No**	3	5	**62.50%**
10	**Yes**	8	0	**100%**
11	**Yes**	8	0	**100%**
12	**Yes**	8	0	**100%**
13	**Yes**	8	0	**100%**
14	**No**	1	7	**87.50%**
15	**No**	0	8	**100%**
16	**Yes**	8	0	**100%**
17	**Yes**	8	0	**100%**
18	**Yes**	8	0	**100%**
19	**Yes**	8	0	**100%**
20	**Yes**	8	0	**100%**
21	**No**	4	4	**50%**
22	**No**	2	6	**75%**
Average IAA rate				**89%**

Table 4.4 – Inter-Annotator Agreement rate on the presence/absence of errors in relation to Perfect Matches (file 002)

Following the same procedure, the IAA rate was also calculated for file 012, where the average agreement rate estimated was around 88%, as shown in the Table 4.5 below.

Segment Unit	**Perfect Match (reference)**	# of annotators spotting mistake in segment		**Agreement with reference (%)**
		Yes	**No**	
1	**No**	1	7	**87.50%**
2	**No**	4	4	**50%**
3	**No**	0	8	**100%**
4	**No**	0	8	**100%**
5	**Yes**	8	0	**100%**
6	**Yes**	8	0	**100%**
7	**Yes**	8	0	**100%**
8	**No**	1	7	**87.50%**
9	**No**	0	8	**100%**
10	**No**	1	7	**87.50%**
11	**No**	2	6	**75%**
12	**No**	2	6	**75%**

13	**Yes**	8	0	**100%**
14	**No**	0	8	**100%**
15	**No**	1	7	**87.50%**
16	**No**	0	8	**100%**
17	**No**	0	8	**100%**
18	**Yes**	8	0	**100%**
19	**Yes**	8	0	**100%**
20	**Yes**	8	0	**100%**
21	**Yes**	8	0	**100%**
22	**Yes**	8	0	**100%**
23	**Yes**	8	0	**100%**
24	**No**	0	8	**100%**
25	**No**	3	5	**62.50%**
Average IAA rate				**92.50%**

Table 4.5 – Inter-Annotator Agreement rate on the presence/absence of errors in relation to Perfect Matches (file 012).

After this first evaluation, the post-test evaluation phase included the calculation of the IAA rate for the Coarse-Grained Error categories (or Layer 1 of the taxonomic scheme described in §3.6). The evaluation aimed therefore at assessing whether all annotators agreed or not on the use of a specific category in the annotation of each segment unit. In the first instance, the agreement rate for the three predefined categories (Deletion, Substitution and Insertion) was calculated for file 002 by also including the "*Null*" category (no entry by part of the annotator) as an additional category to be selected. So, for example, if in a given segment 2 annotators report a "*Null*" error (i.e., they entered none of the above three categories because they did not identify the error or they did not consider it intentionally) and 6 annotators report a Deletion error, the agreement rate for the segment would be 75%. Again, if in a given segment 3 annotators report a Substitution error, 2 annotators report a Deletion error and, finally, 3 annotators report a "*Null*" error, the agreement rate would be 37.5%. It is also important to make it clear that, in the analysis, the calculation of the agreement rate was always done by taking into consideration the predominant category in percentage value (in the example above, the *Null* or Substitution error are the highest categories in percentage). The relevant IAA rate for this first instance is broken down in Table 4.6 below. For a better understanding of the table below, it should be added that, starting from the left, under the column *"# of annotators spotting mistake in segment"*, the number of annotators reporting a mistake under the NUL ("Null"), DEL

("Deletion"), SUB ("Substitution") or INS ("Insertion") categories is specified. Finally, the column on the right of Table 4.6 reports the agreement rate per each segment for the highest score category, as explained above. At the bottom of the Table, the average IAA rate is so calculated for the entire file.

Segment Unit	# of annotators spotting mistake in segment				Predominant category	Agreement (%)
	NUL	DEL	SUB	INS		
1	8	0	0	0	**NUL**	**100%**
2	8	0	0	0	**NUL**	**100%**
3	4	0	4	0	**NUL/SUB**	**50%**
4	8	0	0	0	**NUL**	**100%**
5	0	8	0	0	**DEL**	**100%**
6	4	4	0	0	**NUL/DEL**	**50%**
7	1	7	0	0	**DEL**	**87.50%**
8	8	0	0	0	**NUL**	**100%**
9	3	0	5	0	**SUB**	**62.50%**
10	8	0	0	0	**NUL**	**100%**
11	8	0	0	0	**NUL**	**100%**
12	8	0	0	0	**NUL**	**100%**
13	8	0	0	0	**NUL**	**100%**
14	1	0	7	0	**SUB**	**87.50%**
15	0	0	8	0	**SUB**	**100%**
16	8	0	0	0	**NUL**	**100%**
17	8	0	0	0	**NUL**	**100%**
18	8	0	0	0	**NUL**	**100%**
19	8	0	0	0	**NUL**	**100%**
20	8	0	0	0	**NUL**	**100%**
21	4	0	4	0	**NUL/SUB**	**50%**
22	3	5	0	0	**DEL**	**62.50%**
Average IAA rate						**89%**

Table 4.6 – Inter-Annotator Agreement rate on Coarse-Grained Error category including "Null" errors (file 002)

After this preliminary assessment for Coarse-Grained Error, to better represent the IAA rate among the eight annotators, a further IAA rate was calculated for the same file, but only in relation to the annotators who reported an error, as shown in Table 4.7 below, thus excluding the "null" category from the estimate of IAA rate. This calculation allows seeing and measuring if the reporting annotators chose the same category of error when they identified an error in the segment unit.

Segment Unit	# of annotators choosing the same category when spotting a mistake				**Agreement per category (%)**
		DEL	SUB	INS	
3		0	4	0	**100%**
5		8	0	0	**100%**
6		4	0	0	**100%**
7		7	0	0	**100%**
9		5	0	0	**100%**
14		0	7	0	**100%**
15		0	8	0	**100%**
21		0	4	0	**100%**
22		5	0	0	**100%**
Average IAA rate					**100%**

Table 4.7 – Inter-Annotator Agreement rate on Coarse-Grained Error category only among annotators reporting an error (file 002)

In particular, the average IAA rate reported in Table 4.7 makes it possible to maintain that, when an error is identified in a given segment, the category indication is shared among all annotators having recognized that error. The same method was then used to calculate the IAA rate for file 012. As for the previous file, here the agreement rate on Coarse-Grained Error categories was defined by both including the *"Null"* values and without considering them, as shown in Tables 4.8 and 4.9 below.

Segment Unit	# of annotators spotting mistake in segment				**Predominant category**	**Agreement (%)**
	NUL	DEL	SUB	INS		
1	1	0	7	0	**SUB**	**87.50%**
2	4	4	0	0	**NUL/DEL**	**50%**
3	0	0	8	0	**SUB**	**100%**
4	0	1	7	0	**SUB**	**85.50%**
5	8	0	0	0	**NUL**	**100%**
6	8	0	0	0	**NUL**	**100%**
7	8	0	0	0	**NUL**	**100%**
8	1	0	7	0	**SUB**	**87.50%**
9	0	0	8	0	**SUB**	**100%**
10	1	0	7	0	**SUB**	**87.50%**
11	2	0	0	6	**INS**	**75%**
12	2	0	6	0	**SUB**	**75%**
13	8	0	0	0	**NUL**	**100%**
14	0	0	8	0	**SUB**	**100%**
15	1	0	7	0	**SUB**	**87.50%**

16	0	0	8	0	**SUB**	**100%**
17	0	0	8	0	**SUB**	**100%**
18	8	0	0	0	**NUL**	**100%**
19	8	0	0	0	**NUL**	**100%**
20	8	0	0	0	**NUL**	**100%**
21	8	0	0	0	**NUL**	**100%**
22	8	0	0	0	**NUL**	**100%**
23	8	0	0	0	**NUL**	**100%**
24	0	0	8	0	**SUB**	**100%**
25	3	0	5	0	**SUB**	**62.50%**
Average IAA rate						**92%**

Table 4.8 – Inter-Annotator Agreement rate on Coarse-Grained Error category including "Null" errors (file 012)

Table 4.8 above shows that for file 012 the mean agreement rate was of about 92%, which means that the agreement rate among annotators in choosing one of the four categories (Null, Deletion, Substitution, Insertion) was high. On the other hand, when examining only the segment units where an error was recognized by the majority of annotators (Table 4.9 below), the IAA rate on the selection of the same category was even higher, amounting to 98.30%, approximately.

Segment Unit	# of annotators choosing the same category when spotting a mistake				**Agreement per category (%)**
		DEL	SUB	INS	
1		0	7	0	**100%**
2		4	0	0	**100%**
3		0	8	0	**100%**
4		1	7	0	**87.50%**
8		1	7	0	**87.50%**
9		0	8	0	**100%**
10		0	7	0	**100%**
11		0	0	6	**100%**
12		0	6	0	**100%**
14		0	8	0	**100%**
15		0	7	0	**100%**
16		0	8	0	**100%**
17		0	8	0	**100%**
24		0	8	0	**100%**
25		0	5	0	**100%**

Average IAA rate	**98.30%**

Table 4.9 – Inter-Annotator Agreement rate on Coarse-Grained Error category only among annotators reporting an error (file 012)

The next step of the experimental test involved formulating the Inter-Annotator Agreement rate for the second layer of the annotation data, namely the Fine-Grained Error categories. As explained in §3.6, the Fine-Grained Error categories for the present study are: Lexis, Grammar, Terminology, Disfluency and Prosody. For this IAA rate calculation, it is first necessary to underline that the analysis was expected to reach a lower level of agreement on Fine-Grained Error categories as their number is higher if compared to the previous taxonomic layer (where 3 categories are set up). However, as explained in §4.9 below, a lower IAA rate was also expected because of the criticalities regarding the similarity and potential ambiguity between the category pairs *"Lexis/Terminology"* and *"Lexis/Grammar"* (as also discussed in §3.6.). The decision of keeping these categories separated is based on the fact that the present thesis intends to examine and describe the role of specific domain terminology in the final output accuracy. Additionally, it should be remarked that, in the calculations under this taxonomic layer, when in a given segment unit there were 3 or more category entries (according to the different annotators), the IAA rate was calculated on the prevailing (in quantitative terms) category only. Therefore, if in a given segment unit 4 annotators reported a Grammar category, 1 annotator reported a Lexis category and 3 annotators entered a "Null" category, the agreement rate was calculated on the basis of the prevailing category input: in the example, the Grammar category. In particular, the formula used for this calculation is as follows:

Formula: $N : T = x : 100$

where *"N"* is the number of category inputs (for the prevailing one), *"T"* is the total number of annotators, and *"x"* is the IAA rate to be obtained. For the example above, the IAA rate formula calculation is: $4 : 8 = x : 100$; and the IAA rate is: $x = 50\%$. As already seen for the previous taxonomic categories, also in the case of the Fine-Grained Error categories, the calculations were broken down by including or excluding the *"Null"* values in the estimates. Table 4.10 below shows the IAA rate calculated by including the *"Null"* values (thus considering the entries where annotators did not recognized any error).

Segment Unit	# of annotators spotting mistake in segment						Predominant category	Agreement (%)
	NUL	LEX	GRA	TER	DIS	PRO		
1	8	0	0	0	0	0	**NUL**	**100%**
2	8	0	0	0	0	0	**NUL**	**100%**
3	4	2	2	0	0	0	**NUL**	**50%**
4	8	0	0	0	0	0	**NUL**	**100%**
5	0	0	0	0	8	0	**DIS**	**100%**
6	4	0	4	0	0	0	**GRA**	**50%**
7	1	0	0	0	7	0	**DIS**	**87.50%**
8	8	0	0	0	0	0	**NUL**	**100%**
9	3	0	5	0	0	0	**GRA**	**62.50%**
10	8	0	0	0	0	0	**NUL**	**100%**
11	8	0	0	0	0	0	**NUL**	**100%**
12	8	0	0	0	0	0	**NUL**	**100%**
13	8	0	0	0	0	0	**NUL**	**100%**
14	1	1	6	0	0	0	**GRA**	**75%**
15	0	5	0	3	0	0	**LEX**	**62.50%**
16	8	0	0	0	0	0	**NUL**	**100%**
17	8	0	0	0	0	0	**NUL**	**100%**
18	8	0	0	0	0	0	**NUL**	**100%**
19	8	0	0	0	0	0	**NUL**	**100%**
20	8	0	0	0	0	0	**NUL**	**100%**
21	4	1	3	0	0	0	**NUL**	**50%**
22	3	0	5	0	0	0	**GRA**	**62.50%**
Average IAA rate								**86.30%**

Table 4.10 – Inter-Annotator Agreement rate on Fine-Grained Error category including "Null" errors (file 002)

To better understand the results in Table 4.10 above, it should be noted that, starting from the left, the column *"# of annotators spotting mistake in segment"* includes *"NUL,"* which indicates the number of annotators reporting a "Null" error; *"LEX"*, the number of annotators reporting a Lexis errors *"GRA"* shows the number of annotators reporting a Grammar error; *"TER"*, the number of annotators reporting a Terminology error; *"DIS"*, the number of annotators reporting a Disfluency error; and, finally, *"PRO"* indicates the number of annotators reporting a Prosody error. As it is possible to assess from Table 4.10, the Inter-Annotator Agreement rate is equivalent to 86.30% and it is lower if compared to the rate achieved with the previous taxonomic layer. However, it can be considered as quite satisfactorily for this level of the analysis, as these categories are mostly used in the present study for descriptive purposes and not for the quantitative evaluation of accuracy.

Segment Unit	# of annotators choosing the same category when spotting a mistake						**Agreement (%)**
		LEX	GRA	TER	DIS	PRO	
3		2	2	0	0	0	**50%**
5		0	0	0	8	0	**100%**
6		0	4	0	0	0	**100%**
7		0	0	0	7	0	**100%**
9		0	5	0	0	0	**100%**
14		1	6	0	0	0	**85.71%**
15		5	0	3	0	0	**62.50%**
21		1	3	0	0	0	**75%**
22		0	5	0	0	0	**100%**
Average IAA rate							**86%**

Table 4.11 – Inter-Annotator Agreement rate on Fine-Grained Error category excluding "Null" errors (file 002).

Subsequently, after this preliminary calculation, the agreement rate per category was obtained by excluding those segments where no errors were recognized by the annotators and by taking into account only those annotators who indeed reported an error (as done for the Coarse-Grained Error categories above). The relevant values for the calculation of the IAA rate are reported in Table 4.11 above. In this case too, the IAA rate so obtained resulted to be lower than for the previous taxonomic layer's IAA rate, now amounting to 86%.

When analysing the results for the second file of this study's IAA test (file 012), the Inter-Annotator Agreement rate for the Fine-Grained Error categories was broken down and calculated by implementing the same method described above. In Table 4.12 and Table 4.13 below, the relevant rates by including and excluding the *"Null"* values, respectively, are reported. Moreover, both Tables show that, as expected, with file 012 a lower IAA rate was obtained if compared to previous taxonomic layer. This is due to the increased number of options available (hence the increased statistical probability that each annotator enters a different value). Nevertheless, the lower agreement rate is also due to a certain similarity and potential ambiguity between categories (namely, between *Lexis* and *Grammar* or between *Terminology* and *Lexis*), as already mentioned.

Segment Unit	# of annotators spotting mistake in segment						**Predominant category**	**Agreement (%)**
	NUL	LEX	GRA	TER	DIS	PRO		
1	1	2	5	0	0	0	**GRA**	**62.50%**
2	4	0	0	0	4	0	**NUL/DIS**	**50%**
3	0	4	4	0	0	0	**LEX/GRA**	**50%**
4	0	3	5	0	0	0	**GRA**	**62.50%**
5	8	0	0	0	0	0	**NUL**	**100%**
6	8	0	0	0	0	0	**NUL**	**100%**
7	8	0	0	0	0	0	**NUL**	**100%**
8	1	7	0	0	0	0	**LEX**	**87.50%**
9	0	6	0	2	0	0	**LEX**	**75%**
10	1	4	3	0	0	0	**LEX**	**50%**
11	2	2	4	0	0	0	**GRA**	**50%**
12	1	3	4	0	0	0	**GRA**	**50%**
13	8	0	0	0	0	0	**NUL**	**100%**
14	6	0	2	0	0	0	**NUL**	**75%**
15	1	3	4	0	0	0	**GRA**	**50%**
16	0	3	5	0	0	0	**GRA**	**62.50%**
17	0	4	4	0	0	0	**LEX/GRA**	**50%**
18	8	0	0	0	0	0	**NUL**	**100%**
19	8	0	0	0	0	0	**NUL**	**100%**
20	8	0	0	0	0	0	**NUL**	**100%**
21	8	0	0	0	0	0	**NUL**	**100%**
22	8	0	0	0	0	0	**NUL**	**100%**
23	8	0	0	0	0	0	**NUL**	**100%**
24	0	8	0	0	0	0	**LEX**	**100%**
25	3	0	5	0	0	0	**GRA**	**62.50%**
Average IAA rate								**77.50%**

Table 4.12 – Inter-Annotator Agreement rate on Fine-Grained Error category including "Null" errors (file 012).

Segment Unit	# of annotators choosing the same category when spotting a mistake						**Agreement (%)**
		LEX	GRA	TER	DIS	PRO	
1		2	5	0	0	0	**71.40%**
2		0	0	0	4	0	**100%**
3		4	4	0	0	0	**50%**
4		3	5	0	0	0	**62.50%**
8		7	0	0	0	0	**100%**
9		6	0	2	0	0	**75%**
10		4	3	0	0	0	**57.14%**
11		2	4	0	0	0	**66.66%**
12		3	4	0	0	0	**57.14%**

14		0	2	0	0	0	**100%**
15		3	4	0	0	0	**57.14%**
16		3	5	0	0	0	**62.50%**
17		2	2	0	0	0	**50%**
24		8	0	0	0	0	**100%**
25		0	5	0	0	0	**100%**
Average IAA rate							**74%**

Table 4.13 – Inter-Annotator Agreement rate on Fine-Grained Error category excluding "Null" errors (file 012)

However, it should be remarked that, even if the rates achieved for the Fine-Grained Error taxonomy were slightly lower than those reached for the Coarse-Grained Error taxonomy, it is possible to consider these IAA rate values as substantial, by using the adjective "substantial" according to previous studies (Fuoli and Hommerberg, 2015: 334; Gagliardi, 2018: 5): i.e., when the rate is well above the majority of raters/annotators involved. Obviously, by combining the category *"Lexis"* with *"Grammar"* together, it would be possible to obtain a higher IAA rate for the IAA rate of file 012, as shown in Table 4.14 below. More specifically, the values corresponding to *GRA/LEX* are merged under a hypothetical Lexis+Grammar category denominated "*LG"* (the calculation in Table 4.14 were carried out by excluding the "Null" values). However, this simplified approach is rejected in the present study, as it is here considered as more interesting to examine the specific error categories, as already defined in the methodology.

Segment Unit	# of annotators choosing the same category when spotting a mistake						**Agreement (%)**
			LG	TER	DIS	PRO	
1			7	0	0	0	**100%**
2			0	0	4	0	**100%**
3			8	0	0	0	**100%**
4			8	0	0	0	**100%**
8			7	0	0	0	**100%**
9			6	2	0	0	**75%**
10			7	0	0	0	**100%**
11			6	0	0	0	**100%**
12			7	0	0	0	**100%**
14			2	0	0	0	**100%**
15			7	0	0	0	**100%**
16			8	0	0	0	**100%**

17			4	0	0	0	**100%**
24			8	0	0	0	**100%**
25			5	0	0	0	**100%**
Average IAA rate							**98%**

Table 4.14 – Inter-Annotator Agreement rate on Fine-Grained Error category combining the Lexis/Grammar categories (file 012)

Obviously, as it can be seen above, the Inter-Annotator Agreement (IAA) rate would be increased to an average rate of 98% in the example. In this respect, it should be remarked once again that this layer of taxonomy (Layer 2) has little impact on the evaluation of accuracy in quantitative terms and it is not addressed to the calculations of accuracy (namely, the WER and NER indexes) in the automatic software transcriptions generated by *VoxSigma* and Google Speech Recognition engine (via *YouTube* and *Descript*). However, this taxonomic scheme is important for the description of error categorization and it covers a certain relevance in the discussion of results (as better described in §4.9).

Finally, to complete the evaluation of the reliability and transparency of the annotation scheme adopted in the present study, it is necessary to verify the Inter-Annotator Agreement (IAA) rate for the Error Seriousness categories: *i.e.*, "Serious" and "Not Serious". This classification is in fact of absolute importance and significance in the calculations of accuracy, as better defined in §4.4.3 below. Before discussing the Tables below, it should be underlined that for this categorization, the segments with at least 1 error were only taken into account, and that the estimates were done on the basis of the number of annotators who effectively recognized and reported that error. In the first instance, the IAA rate was calculated for file 002, as shown in Table 4.15 below.

Segment Unit	# of annotators reporting a Not Serious Error	# of annotators reporting a Serious Error	Agreement %
3	4	0	**100%**
5	8	0	**100%**
6	2	2	**50%**
7	7	0	**100%**
9	5	0	**100%**
14	4	2	**66.66%**
15	0	8	**100%**

21	2	2	**50%**
22	5	0	**100%**
Average IAA rate			**85%**

Table 4.15 – Inter-Annotator Agreement rate on the category Serious/Not Serious (file 002)

At this point, by following the same procedure, it is possible to carry out the calculation for file 012, by considering only the segments and the annotators reporting and recognizing the error in the given segment, as shown below in Table 4.16 below.

Segment	# of annotators reporting a Not Serious Error	# of annotators reporting a Serious Error	Agreement %
1	1	6	**85.71%**
2	4	0	**100%**
3	1	7	**87.50%**
4	1	7	**87.50%**
8	0	7	**100%**
8	6	0	**100%**
9	0	8	**100%**
10	3	4	**57.14%**
11	6	0	**100%**
12	3	4	**57.14%**
14	2	1	**66.66%**
14	0	8	**100%**
15	1	6	**85.71%**
16	4	4	**50%**
16	2	2	**50%**
17	0	8	**100%**
24	0	8	**100%**
25	5	0	**100%**
Average IAA rate			**84.85%**

Table 4.16 – Inter-Annotator Agreement rate on the category Serious/Not Serious (file 012)

As shown above, the IAA rate was of about 84.02%, which can again be considered as a substantial value for the purposes of the evaluation of the annotators' inputs, compared to previous similar studies (Fuoli and Hommerberg, 2015: 334; Gagliardi, 2018: 5).

To conclude the testing of the taxonomy scheme used in the present study, it is possible to draw some preliminary conclusions on the validity and reliability of the set of categories defined and implemented in the analysis of audio/video files (see sections below). First of all, it is possible to claim that a larger portion of the annotators involved in the test have identified the presence or absence of an error with respect to the given Perfect Matches (PM): 89% for file 002 and 92.5% for file 012. Secondly, a high Inter-Annotator Agreement rate was found for the taxonomic scheme Layer 1 (Coarse-Grained Error categories) of this study, for both files. In particular, for file 002, it was possible to reach a IAA rate of 89% and 100% (including and excluding *"Null"* errors, respectively), while for file 012 the IAA rate was of 92% and 98.30%, respectively. For the Fine-Grained Error taxonomy, the rates were lower and, as already mentioned above, this is mostly due to a major ambiguity between pairs of categories and to the higher probability of entering a different value (as there are 5 different categories to choose from). However, these values of agreement remain substantial and can prove the validity of this taxonomic level too. A separate final consideration should be made about Layer 2 Taxonomy: the lower IAA rate achieved was not only due to the higher number of categories if compared to Layer 1, but also to specific source speech features present in file 012 or due to different interpretation of categories by the single annotators. This aspect should have been examined further by involving the annotators in a post-test phase but this was not possible as the voluntary annotators participating into the analysis were not initially required to take part into a follow-up phase. More specifically, for file 002, it was possible to obtain a IAA rate of 86.30% and 86% (including and excluding "Null" errors, respectively), while for file 012 the IAA rate was of 77.50% and 74%, respectively. Finally, when considering the error severity categorization (the pair *Serious/Not Serious*), the IAA rate achieved a good level of agreement among the 8 annotators, with values of 85% (file 002) and 84.85% (file 012), respectively. These values allow considering this study's taxonomic scheme to be as sufficiently reliable and reproducible, given the substantial levels achieved (to use the conceptual categorization discussed in Fuoli and Hommerberg, 2015: 334; Gagliardi, 2018: 5), especially at the level of Coarse-Grained Errors. The taxonomic scheme so elaborated and defined will be implemented in the next sections of this chapter for the analysis of software automatic transcriptions and the annotation data. Finally, it should be highlighted that, given the limited set of segments which was examined, these results and experimentation have a reduced application and cannot be considered as a 100% test solution for the present and the future studies, though providing substantial basis for its validity. Probably, with the selection of a highly expert pool of annotators in the field of ASR annotation and transcription, the results would have been even more substantial.

4.4. Analysis of Annotation Data from VoxSigma transcriptions

In this section, the study will present and evaluate the results obtained from the annotation process completed after the automatic transcription generated by Vocapia Research's speech recognition solution – *VoxSigma*. In particular, the annotation data will be analysed according to the two different layers described in Chapter 3, §3.6: Layer 1 Taxonomy (Coarse-Grained Error categorization) and Layer 2 Taxonomy (Fine-Grained Error categorization). The evaluation of data will include a series of graphs per each file for a description of the errors distribution, as well as a review of common criticalities and problems in the automatic speech recognition process, both for Native and Non-Native files. To conclude this section, an analysis of the *Serious/Not Serious* category of errors will be discussed, while trying to make a distinction between what is considered as "serious" and thus having an impact on the accuracy of automatic recognition and what is considered as "not serious".

4.4.1. Analysis of Layer 1 Taxonomy Errors

As seen in §3.6, the three major categories of error for Layer 1 Taxonomy (Coarse-Grained Errors) are Deletion, Substitution and Insertion. For the purposes of the present study, it was decided to underline the common criticalities and problems concerning automatic speech recognition and the relevant taxonomic scheme by analysing the data according to two groups of speakers: Native and Non-Native (the main variable). The first evidence of a better accuracy in Native-speaker files emerges from the analysis of the aggregate number of errors (Substitution + Deletion + Insertion) across the database of files. In particular, as shown in Figure 4.5 below, for Native-speaker files, the aggregate number of errors is of 784 errors, while for Non-Native files, it amounts to 1692 errors (more than double the number of errors obtained in Native group).

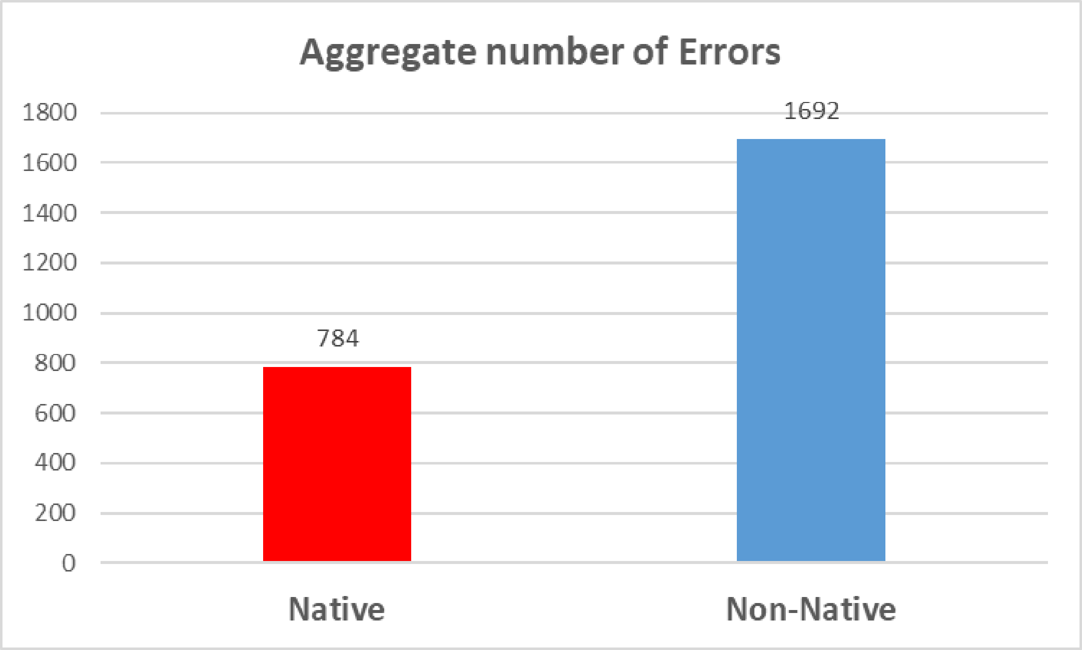


Figure 4.5 – Aggregate number of Errors for Layer 1 Taxonomy

Accordingly, if the total number of Perfect Match (PM) segments is analysed with respect to the total number of segment units, it is possible to have a first glimpse of accuracy in the two different groups. In particular, it is possible to observe that the aggregate number of PMs amounts to 1436 over a total number of 2038 segment units, in the case of Native speaker files, and to 1608 PMs over a total of 2915 segment units in the case of Non-Native speaker files. If these aggregate values are compared to the total number of segments for both groups of files, the following percentage values are obtained: 70.40% for Native and 55.16% for Non-Native speakers. From this very first, rough evaluation of accuracy based on Perfect Matches, it is thus possible to discover that for Native speaker files the ASR technology provides for better results, if compared to the Non-Native group, as shown in Figure 4.6 below: PM indicates the percentage of Perfect Match segments, while FM indicates the percentage of Fuzzy Match segments.

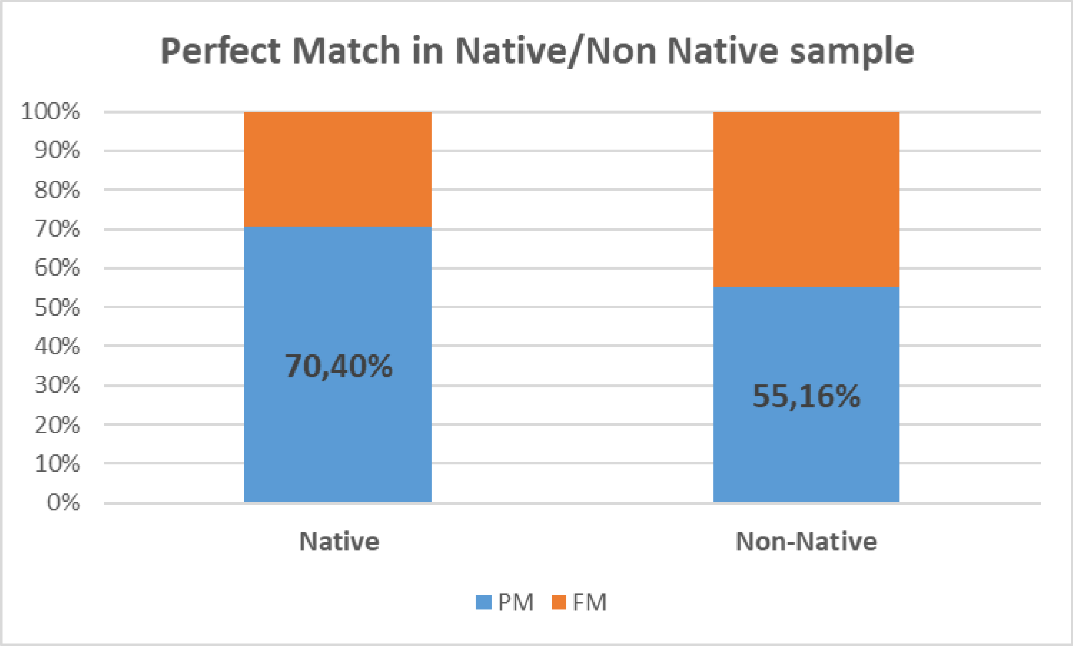


Figure 4.6 - Number of Perfect Match (PM) over the total segment number.

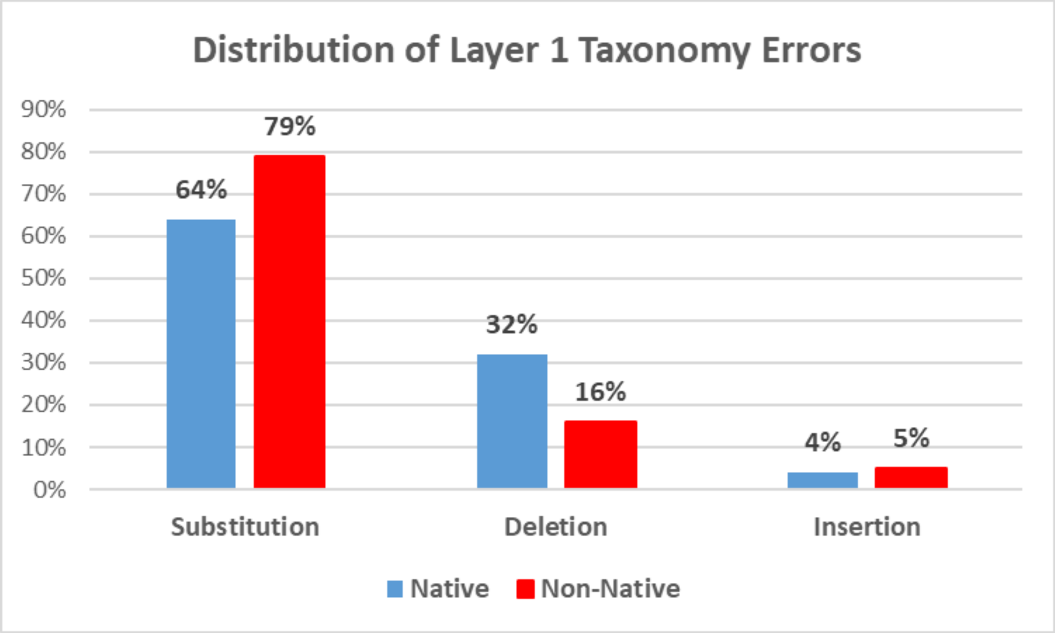


Figure 4.7 - Distribution of Layer 1 Taxonomy errors for Non-Native/Native groups.

More specifically, for the Non-Native group, Figure 4.7 above shows that, across the sample of files, the Substitution errors are the predominant (1337 Errors; 79%) category, while the Deletion (272 Errors; 16%) and Insertion (83 errors; 5%) errors cover a significantly lower share. In Figure 4.7, a similar, comparable distribution seems to be replicated across the Native group of files as well, where the Coarse-Grained Errors are distributed as follows: 506 errors for the Substitution category, 249 for Deletion and 29 for Insertion. Additionally, it is interesting to observe the greater percentage of Deletion occurrences in Native speakers if compared to the Non-Native speakers, along with a higher percentage of Substitutions in Non-Native speakers if compared to Native speakers. To continue with the analysis, if the two groups of files are compared, it is immediately clear that, for Native speaker-based files, a lower total number of errors for each category is given, in line with previous analysis of Perfect Match values. For the purposes of the accuracy evaluation, this piece of data may represent an early indication of a better automatic recognition process occurring with Native speakers. It is also possible to claim that the distribution highlighted here is similar to that achieved in other ASR projects or studies (for example, see the works by Eugeni, 2009; Dumouchel et al., 2011; Romero-Fresco, 2011; Romero-Fresco and Martínez, 2015), where the ASR process generated a higher portion of Substitution errors with respect to the Deletion and the Insertion categories. As described in Errattahi et al. (2016), this is partially due to the fact that an ASR engine tends to substitute or replace a word or series of words with another word or series of words when the original ones cannot be easily recognized, without leaving empty fragments in the text. Although this phenomenon may happen at a lower frequency rate, deletions (also called "omissions" in the scientific literature) generally occur when the software cannot recognize the total sound of a word or expression due to several motivations: for example, when the speaker pronounces the word in a wrong way (with respect to the language model of the ASR) or in a regional variety, or he/she speaks too rapidly, or even when the word is domain specific and it is not included into the ASR system's vocabulary. Another common behaviour of ASR software reported in the literature (also confirmed in the present analysis of the *VoxSigma* output) is the application of Insertions when the software attempts to correct the original source text according to the grammar or syntactic structure of the sentence. In some occasions, this phenomenon occurs when the speaker's speed rate is so slow that the software completes the sentence without waiting for the completion of the utterance.

After these general considerations, to further substantiate this analysis, a series of examples for each error category is now presented. Additionally, it should be remarked that practical examples of errors are also provided in §4.9 dedicated to the Discussion of results. The examples reported below and their discussion are essential to substantiate the current analysis and accuracy evaluation methodology. It should also be highlighted that the ASR technology output is examined here for the purposes of the identification of its main features and for the validation of this study's taxonomic scheme solidity (in addition to the inter-annotator agreement method). As indicated in Errattahi (2018: 1), *"the key problem in ASR error detection is the identification of effective features"*. As better described in §4.9, several studies in the reference scientific literature tried to underline the main features of ASR by using a varied and often confused categorization of errors.

By starting from the most frequent error typology in this study, Substitution, it should be recalled that the analysis reports this kind of error when the ASR software replaces a word or a series of words with another word or series of words. For example, in the Native-speaker file 012, it is possible to identify a Substitution error in segment 24, as shown in Table 4.17 below.

Segment	**Timestamp**	**Reference Speech**	**ASR output**
21	00:01:35,610 --> 00:01:39,090	*for our people. We want to make sure that our people have safe*	for our people we want to make sure that our people have safe
22	00:01:39,090 --> 00:01:43,720	*food and affordable food and we don't want them to have to*	food and affordable food and we don't want them to have to
23	00:01:44,580 --> 00:01:46,790	*be in a position that they can't*	be in a position that they can't
24	00:01:47,310 --> 00:01:50,970	*deal with shocks when you know, like when hurricanes*	deal with sharks when you know like when hurricanes
25	00:01:51,510 --> 00:01:54,900	*and that sort of things happens. Okay, great.*	and that sort of thing happens. Okay, great.

Table 4.17 – Substitution error in Native-speaker ASR output (extract from file 012)

Here the ASR software, VoxSigma did not correctly recognize the term "shocks" and replaced it with the term "sharks", compromising the meaning of the entire segment unit. This particular error may be due to different reasons that cannot be fully identified, probably depending on several factors such as an error of the software decoder feature or the high number of words (density) uttered by the Native speaker in that given segment. Given the limited data available, it is not possible to identify the causes for this error. However, for a further categorization of this error (as better explained in §4.4.2 below), it is possible to consider it as a Lexis error under the taxonomic scheme (see §3.6) and with a Serious grade. In fact, this error significantly contributes to limit the final user in understanding the meaning of the segment and the entire part of the speech. The second example of Substitution is taken from a Non-Native speaker discourse, file 023, where the term "FAO" is replaced by adjective "foul", as shown in Table 4.18 below.

Segment	Timestamp	Reference Speech	ASR output
1	00:00:11,740 --> 00:00:13,750	*I really want to thank FAO and*	I really want to thank foul and
2	00:00:14,360 --> 00:00:17,090	*Dr. Kundavi who is my direct counterpart in Bangkok*	Dr. can be who is my direct counterpart in Bangkok
3	00:00:17,740 --> 00:00:19,450	*for the invitation, and of course*	for the invitation. And of course
4	00:00:20,140 --> 00:00:24,520	*the hosting by the Fijian government is greatly appreciated.*	the hosting by the Fijian government is greatly appreciated.

Table 4.18 - Substitution error in Non Native-speaker ASR output (extract from file 023)

The error in the example was probably due to the mispronunciation (with respect to the ASR language model) by the Non-Native speaker for the "FAO" abbreviation, which was pronounced by the Non-Native speaker as */faʊl/* (and not as per standard English pronunciation for the abbreviation: /ef/-/ei/-/o/), thus generating an error by the decoder component. However, this recognition error may also be due to the terminological resources incorporated into the ASR software vocabulary. In fact, even if VoxSigma responds to the LVCSR requisite defined in §2.2.3.2, there is a high probability that this term was not included into the decoder vocabulary. As it is not possible to examine the software decoder feature (i.e., the built-in vocabulary is not known to users of the software), the cause of this error cannot be ascertained for sure. In segment 2 of the same file, another Substitution error is identifiable where the proper name "Dr. Kundavi" was replaced by the series of words "Dr. can be". Here the error was probably due to the fact that the proper name is not included into the software vocabulary, but again this was not verifiable at the decoder level. Together with specific terminology terms or domain-related words, proper names represent a large portion of all terminology-related errors, probably due to the fact that they are not included in the decoder vocabulary or to the fact that their pronunciation represents a challenge for the ASR system. For both errors, the categorization would be "Terminology" according to the Layer 2 Taxonomy and "Serious" for the error severity grading. To complete the examples of Substitution errors, it is possible to examine file 045, a speech held at the European Parliament by a Non-Native member of Parliament. Here it is likely that the high phonological density of the discourse is the cause of several Substitution errors. With high phonological density, the present study is here referring to the concept of neighbourhood density. The neighbourhood activation feature is based on the number of phonologically similar words in the lexicon (Luce and Pisoni, 1998). However, other reasons are probably determining the errors below in Table 4.19, for example they may coincide with a mispronunciation of the words as per the standard English version of the ASR system (the speaker is non-native) or even the high pitch of the speaker (average pitch around 143 Hz). Extremely high pitch or low pitch speeches are often associated with a higher number of errors as examined in Liu et al. (2019).

Segment	Timestamp	Reference Speech	ASR output
12	00:00:47,370 --> 00:00:50,070	*but today it's a social movement,*	but today it's a social movement
13	00:00:50,580 --> 00:00:55,860	*hundreds of thousands in more than 70 countries, in more than*	hundreds of thousands in more than 70 countries in more than
14	00:00:55,860 --> 00:01:00,910	*700 cities are going to manifestations*	700 cities going to manage stations
15	00:01:01,920 --> 00:01:05,880	*on this movement, Fridays for future.*	on this movement Friday's for future.
16	00:01:05,880 --> 00:01:11,010	*This is a social revolution that started now*	This is a social revolution that started now
17	00:01:11,190 --> 00:01:17,190	*everywhere, and - dear colleagues - we the politicians, we the parliamentarians,*	everywhere and dear colleagues be the politicians we the parliamentarians.
18	00:01:17,190 --> 00:01:19,650	*We have to be part of that.*	We have to be part of that.
19	00:01:19,650 --> 00:01:25,800	*This is my firm conviction, and one element more.*	This is my firm conviction and one element mall.

Table 4.19 - Substitution errors in Non Native-speaker ASR output (extract from file 045)

More specifically, it is possible to see in segment 14 that the term “manifestations” is replaced with the words, “manage stations”: these terms may be considered as phonetically similar, if not near-homophones, to the source term, and the Lexis category replacement actually determines a Serious error according to the validated taxonomy. Again, in segment 17, the pronoun “we” (used as a speech marker) is replaced with the verb “be”, which is again phonetically similar to the pronoun and not recognized by the system. Pronouns and articles are often misrecognized by the ASR software given their short, phonetic pronunciation. However, in this case, this Lexis error was graded as Not Serious because the segment unit remains fully understandable to the final users, by also considering what comes in the previous and following segments. On the contrary, the replacement of the adverb “more” in segment 19 with the word “mall” (phonetically similar) was considered as a Serious error (Grammar category for Layer 2 Taxonomy) as it compromises the meaning of the segment unit.

Finally, in file 048, a series of Substitution errors occurred probably because of the high-density discourse and because of the high speed rate (155.71 wpm) of the Native speaker. In all Grammar-category examples shown in red below in Table 4.20, Substitution errors are not of serious entity but they may partially compromise the understanding of the relevant single segment units. However, given that the meaning of the entire speech section from 73 to 79 is clearly understandable as a whole, they were accounted for Not Serious errors, expect in the first case where the pronoun “it” was replaced with “he”: the error was considered as a Serious error because the use of the pronoun “he” seems to make reference in the ASR output to an unknown third party. In segment 79, it is also possible to observe a typical example of Grammar error (Not Serious), a phenomenon which is very common across the present study’s ASR output: that is to say, the modification of the verb form (like in this case, where the “-s” of the third person is omitted in the verb “hold”) or of the verb tense (for example, in “has/had”; “rise/rose”; “get/got”, etc.).

Segment	**Timestamp**	**Reference Speech**	**ASR output**
73	00:04:41,380 --> 00:04:45,210	*Now it is sometimes suggested that a belief in a free-market economy*	Now he just sometimes suggested that a belief in a free market economy

74	00:04:45,320 --> 00:04:49,970	*which pursues the objective of economic growth is not compatible with taking*	which pursues the objective of economic growth is not compatible with taking
75	00:04:49,970 --> 00:04:54,290	*the action necessary to protect and enhance our natural environment.*	the action necessary to protect and enhance our natural environment.
76	00:04:54,290 --> 00:04:56,750	*Then we do need to give up on the very idea of economic growth*	Do we need to give up on the very idea of economic growth
77	00:04:56,750 --> 00:05:01,050	*itself as the price we have to pay for sustainability.*	itself as the price we have to pay for sustainability.
78	00:05:01,050 --> 00:05:04,640	*Others argue that taking any action to protect and improve our environment*	Others argue that taking any action to protect and improve our environment
79	00:05:04,710 --> 00:05:07,700	*harms business and holds back growth.*	harms business and hold back growth.

Table 4.20 - Substitution errors in Native-speaker ASR output (extract from file 048)

For a few examples of the Deletion error category, it is now possible to review the file 049, as shown in Table 4.21 below, where a couple of ASR deletions are reported under the Grammar category. Here, probably due to the Native-speaker's medium speed rate (136 words/min) and the phonological density of discourse, the ASR system did not recognize article "the" (in segment 119) and verb "are" (in segment 121). However, both errors were considered as Not Serious, as they did not alter the meaning of the discourse portion in question, nor they changed the meaning of the single segment units in which they occurred.

Segment	**Timestamp**	**Reference Speech**	**ASR output**
117	00:08:22,850 --> 00:08:27,710	*Children should not have to pay with their health for our failure to*	Children should not have to pay with their health for our failure to
118	00:08:27,780 --> 00:08:32,030	*clean up our toxic air, in a moment.*	clean up our toxic air in a moment,
119	00:08:32,030 --> 00:08:36,740	*And it's the working class communities that suffer the worst effects of air*	and it's (the) working class communities that suffer the worst effects of air
120	00:08:36,880 --> 00:08:42,100	*pollution, those who are least able to rebuild their lives after flooding,*	pollution. Those who are least able to rebuild their lives after flooding
121	00:08:42,630 --> 00:08:46,690	*will be hit hardest by rising food prices while the better off who are*	will be hit hardest by rising food prices while the better off who (are)

122	00:08:46,860 --> 00:08:48,950	*sometimes more responsible for the most*	sometimes more responsible for the most
123	00:08:49,370 --> 00:08:51,840	*emissions can pay their way out of the trouble.*	emissions can pay their way out of the trouble
124	00:08:52,400 --> 00:08:56,180	*And internationally, in a cruel twist of fate,*	and internationally in a cruel twist of fate,

Table 4.21 - Deletion errors in Native-speaker ASR output (extract from file 049)

In file 030 the Non-Native speaker uttered the word "food" thrice and seemed to be hesitant when speaking, therefore determining a Deletion error by part of the ASR system, as shown in Table 4.22 below, segment 22. However, it is not possible to claim that the error was due to the speaker's hesitation or false start, or due to the Non-Native pronunciation of the term "food" itself. Although the word appears after the sentence verb, provided that the omitted occurrence of the word "food" is coincident with the subject of the sentence, this error was considered as a Serious error as it may compromise the single segment understanding.

Segment	**Timestamp**	**Reference Speech**	**ASR output**
22	00:01:33,830 --> 00:01:37,820	*For a long time food has been looked upon, the food and the food*	For a long time (food) has been looked upon the food and the food
23	00:01:37,820 --> 00:01:43,160	*supply, has been looked upon as a central function of the central government*	supply had been looked upon as a central function of the central government
24	00:01:43,370 --> 00:01:47,000	*of Sri Lanka, but now we can see more and more the*	of Sri Lanka, but now we can see more and more the
25	00:01:47,090 --> 00:01:51,410	*provincial government and also the local government coming into the scene,*	provincial government and also the local government coming into the scene

Table 4.22 - Deletion error in Non Native-speaker ASR output (extract from file 030)

For a few examples of Insertion errors, it is possible to survey files 037 and 055, as shown in Tables 4.23 and 4.24 below. In the first of the two files, in segment 38, the conjunction "and" is added by the ASR system (not present in the reference speech), probably because of the high-density list of terms in the speech, or because of the mispronunciation by the Non-Native speaker (with respect to the ASR language model) in the entire segment. However, it is more likely that the confusion made by VoxSigma software was generated by the assonance and combination of the end letters of the term "climate" and the start letters of the term "energy". Alternatively, this mistake may also be justified by the potential presence of a disfluency element, for example the speech filler "um" which could have been recognized as the sound "and" by the software (but its presence is not verifiable by attentively listening to the speech). Therefore, when analysing these data, it should be remarked that several factors enter into play into the production of errors by ASR software. Hence the necessity of adopting a simple taxonomic scheme without trying to justify or clarify the errors by using the conventionally-adopted features generally implemented in the scientific literature, as better explained in the Discussion of results (§4.9). To my judgement, most ASR errors are due to multiple reasons that cannot be clarified by using a single feature or categorization of error. In the same file, an example of Substitution is also available: the word "investment" is replaced with "in the west men" words.

Segment	Timestamp	Reference Speech	ASR output
37	00:02:25,330 --> 00:02:28,480	*Under this Commission we have learned how to better integrate*	under this Commission we have learned how to better integrate,
38	00:02:28,480 --> 00:02:33,230	*climate, energy, transport and other policies into the Energy Union.*	climate and energy, transport and other policies into the energy union
39	00:02:33,860 --> 00:02:38,330	*And we are the world leader in designing coherent policies that drive investment*	and we are the world leader in designing coherent policies that drive in the west men

Table 4.23 - Insertion error in Non Native-speaker ASR output (extract from file 037)

This is a Serious, Lexis error probably due to the mispronunciation by the speaker (with respect to the ASR language model), who separated the syllables "invest-" from the syllable "-ment" with a pause. Finally, in order to conclude with the presentation of some examples of Insertion errors (Coarse-Grained Error taxonomy), in file 055, segment 1, the pronoun "I" is automatically, wrongly added by the ASR system probably because the grammar code system (language model) incorporated into

VoxSigma can hardly recognize the use of the imperative form of the verb "think", typical of a political speech or discourse. This Grammar error could thus be explained as an error due to the misinterpretation of intonation.

Segment	Timestamp	Reference Speech	ASR output
39	00:02:44,990 --> 00:02:47,580	Now think about the shame that each of us will carry when	Now I think about the shame that each of us will carry when
40	00:02:47,580 --> 00:02:51,360	our children and grandchildren look back and realize that we had the means	our children and grandchildren look back and realize that we have the means
41	00:02:51,510 --> 00:02:56,460	of stopping this devastation, but simply lacked the political will to do so.	of stopping this devastation but simply lacked the political will to do so.

Table 4.24 - Insertion error in Native-speaker ASR output (extract from file 055)

Further examples of Layer 1 Taxonomy (Coarse-Grained Errors) can be found in §4.9, which is dedicated to the Discussion of results, together with a comparison with the reference literature's most common features used to describe and categorize them. At this point, the evaluation of annotation data is now carried out for the Fine-Grained Error categorization.

4.4.2. Analysis of Layer 2 Taxonomy Errors

For the analysis and evaluation of Fine-Grained Error categories (Layer 2 Taxonomy, as described in Chapter 3 on Methodology, §3.6), it is important to clarify that the five categories so defined (*Lexis, Grammar, Terminology, Prosody* and *Disfluency*) are examined as a further statistical measure of errors distribution, contributing in minimum terms to the evaluation of ASR process accuracy. In particular, the weight of certain categories with respect to others, as well as their distribution across Native and Non-Native files is examined. Yet it should be remarked, as already specified in Chapter 3, that this set of categories is primarily annotated and used for descriptive purposes, and that the relevant annotation data do not contribute to the calculation of accuracy according to the WER and NER models.

By starting the analysis with Non-Native speaker files, it is possible to observe that the distribution of errors is as follows: *Grammar*, 786 (46%); *Lexis*, 523 errors (31%); *Terminology*, 271 (16%); *Disfluency*, 104 errors (6%); and, finally, *Prosody*, with just 8 errors (1%). The graphic representation of the distribution is offered in Figure 4.8 below. An almost equivalent distribution can be observed in Native files too. More specifically, the ASR process generated 336 errors for *Grammar* (43%), 163 errors under the *Lexis* (21%) category, 132 errors for *Terminology* (17%), 138 errors for *Disfluency* (17%) and, finally, 15 errors for the *Prosody* category (2%). Both in the Non-Native files and in Native groups of files, the *Lexis* and *Grammar* categories account for the majority of error occurrences: 77% and 64%, respectively.

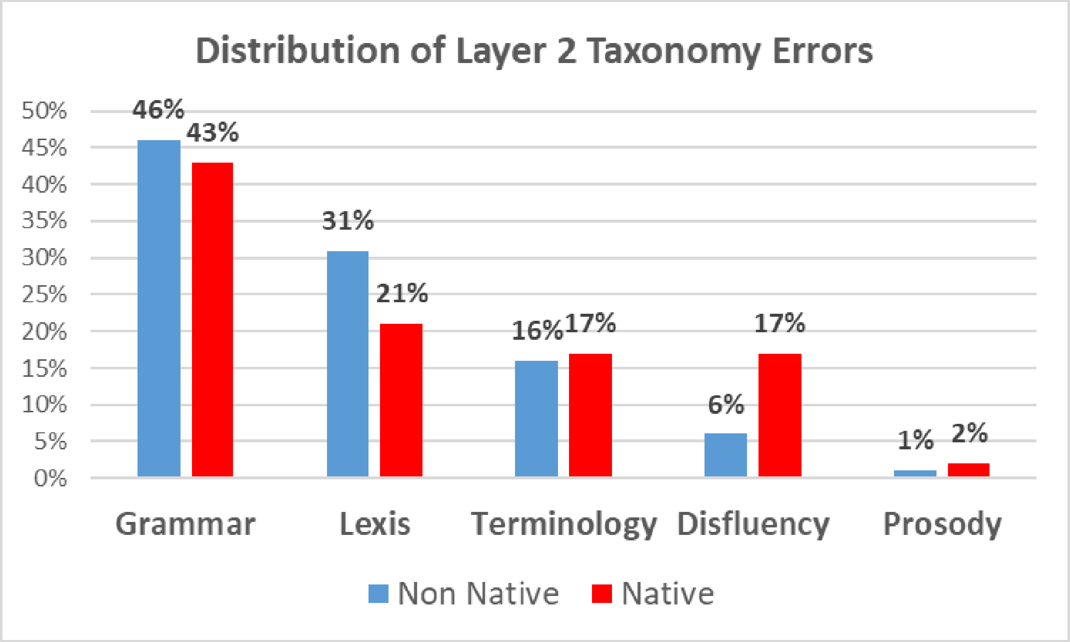


Figure 4.8 – Distribution of Fine-Grained Errors in Non-Native/Native speaker files.

In both groups of files, Grammar and Lexis categories represent the major share of errors: 77% in Non-Native speaker files and 67% in Native speaker files. Furthermore, it is interesting to observe, at this stage, that the Disfluency-based errors are significantly higher in percentage (17%) in Native files (if compared to Non-Native ones, 6%). Finally, it is observable that Terminology-based errors account for an almost equivalent share 16% in Non-Native and 17% in Native). This piece of data, together with the Lexis component, is particularly interesting to be examined (as better exemplified in the Discussion of results: §4.9) as it may represent a useful basis for the enhancement and optimization of the ASR system's Augmented Terminology feature. In order to substantiate the analysis of the Fine-Grained Error taxonomic scheme further, examples of errors are now offered here below.

To start with Grammar errors that occur when the ASR system does not recognize a grammar rule properly or correctly (e.g., verb tense, verb form, prepositions in phrasal verbs, adverbs, etc.), it is possible to survey Native-speaker file 010, as shown in Table 4.25 below. In this extract, the adverb "quite" was replaced in segment 21 by the adjective "quiet", which has the same phonetic sound: */kwaɪt/* (they could be considered as "homophones"). Although grammatically this error may appear as a serious one, actually the final user of the subtitle unit could understand the meaning of the segment unit: the Substitution error in question was therefore graded as Not Serious. However, if the previous error in segment 21 is examined, the entire segment unit becomes significantly hard to be understood. In fact, the omission of the verb "scrumbled" (Deletion, Serious error) poses a serious challenge for the subtitle viewer/reader's comprehension. In the file, an example of Disfluency error is also present in segment 20, where speech filler "uhm" is deleted by the ASR system (Deletion error, Not Serious grading).

Segment	Timestamp	Reference Speech	ASR output
20	00:01:20,490 --> 00:01:24,300	*Uhm. The World Bank's Global Partnership for the Ocean,*	[uhm] The World Bank's Global Partnership for the ocean,
21	00:01:24,300 --> 00:01:25,410	*scrumbled a little bit quite recently,*	[scrumbled] the little bit quiet recently,
21	00:01:24,300 --> 00:01:25,411	*scrumbled a little bit quite recently,*	[scrumbled] the little bit quiet recently,

Table 4.25 – Grammar error in Native-speaker ASR output (extract from file 010)

By taking into consideration Non-Native speaker file 007, it is possible to examine a series of three different Grammar errors in segments 1, 3 and 5 (see Table 4.26 below). More specifically, in segment 1, the omission of the subject of the sentence "I" (pronoun) is considered a Serious error, while in segment 3, the deletion of preposition "about", plus article "a", represents a minor error. These errors are generally due to mispronunciation by the speaker with respect to the correct pronunciation of the word in English (ASR language model) or the high phonological density of speech. More interesting is the example in segment 5, where it is possible to find a typical substitution of the verb tense in the verb "discussed" (replaced with the present tense of the verb). Many of the Grammar errors do in fact refer to verb tense or verb form changes as they are very similar in terms of pronunciation (actually, they could be considered as near-homophones) and thus they are difficult to be recognized by the ASR system in high phonological density discourse. However, other factors may include the mispronunciation by Non-Native speakers (with respect to the ASR language model), who often tend to omit the "-s" in singular (as per the correct pronunciation in English), third person form of verbs or the "-ed" ending in the simple past forms of the verb.

Segment	Timestamp	Reference Speech	ASR output
1	00:00:04,030 --> 00:00:06,790	*I was always pleasant to discuss and uhm*	(I) Was always pleasant to discuss and [uhm]
2	00:00:09,600 --> 00:00:12,420	*have a conversation about many things,*	have a conversation about many things,
3	00:00:12,420 --> 00:00:16,500	*especially about a concern that we have discussed before*	especially [about a] concern that we have discussed before,
4	00:00:16,500 --> 00:00:19,030	*we met in New York and*	we met in New York and
5	00:00:20,010 --> 00:00:22,410	*we discussed a lot about illegal fishing*	we discuss a lot about the illegal fishing

Table 4.26 – Grammar error in Non-Native ASR output (extract from file 007)

To find errors for the Lexis category, the analysis should consider, for example, Non-Native speaker file 014 (see Table 4.27 below). As seen in §3.6, Lexis errors are those errors where the ASR system fails in recognizing (or even add) lexical elements or nouns correctly into the speech, if compared to the gold standard transcription. In the extract below, the ASR system failed to recognize the term "international donors" in segments 2-3 (replaced with "of the year I do"). Under this specific case, the speaker fluency in English turned out to be very difficult to be interpreted by the ASR system (and also by a human listener during the manual transcription of speech): the term “donors” was wrongly pronounced by the Non-Native speaker as */dʊnə/* instead of using the correct version: UK: *ˈdəʊnə*^{r}*/* US:*/ˈdoʊnɚ/ ,(dōˈnər)*. On the contrary, in segment 1, the Substitution of the term “FAO” with adjective “foul” should be ascribed to wrong recognition of specialized or domain-related terminological resources, and thus be considered as a Terminology error. In both cases, the severity of these errors was graded as Serious because both segment units resulted to be incomprehensible.

Segment	**Timestamp**	**Reference Speech**	**ASR output**
1	00:00:13,170 --> 00:00:16,380	*Well, first thing FAO was the first*	Well, first thing foul was the first
2	00:00:16,890 --> 00:00:19,220	*of all international*	of all of the year.
3	00:00:19,220 --> 00:00:22,720	*donors to make a direct contribution*	I do to make a direct contribution

Table 4.27 – Lexis error in Non-Native ASR output (extract from file 014)

Another example of Lexis error is evident in Native-speaker file 016, an extract of which is reported in Table 4.28 below.

Segment	**Timestamp**	**Reference Speech**	**ASR output**
6	00:00:21,200 --> 00:00:23,892	*I just wanted to to, uhm*	I just want to do [uhm]
7	00:00:23,890 --> 00:00:27,340	*think more about what it's like*	I think more about what it's like
8	00:00:28,610 --> 00:00:29,850	*to begin as a seed*	to begin as a seat.
9	00:00:31,720 --> 00:00:33,550	*and to end as a forest*	And to end as a forest
10	00:00:34,600 --> 00:00:37,750	*and that has to find a place to germinate.*	and that as to find a place to germinate,

Table 28 – Lexis error in Native-speaker ASR output (extract from file 016)

In this speech, the ASR system failed to recognize the term “seed” in segment 8, by replacing it with the word “seat”. As the discourse is here made by a Native speaker (and there was not an occurrence of mispronunciation), this Substitution error may tentatively be explained as a weakness of the ASR system, depending on the decoder or on the phonological neighbourhood phenomenon. Lexical problems are often associated throughout the study's output to homophone or near-homophone words, which represent a serious challenge for the ASR process.

Errors belonging to the Terminology category should not be confused with Lexis errors, though a certain ambiguity may arise under specific circumstances, as underlined in the Discussion of results (§4.9 below). According to the taxonomy implemented in this study, Terminology errors generally occur when the ASR system does not recognize correctly, or deletes, domain-related or specialized terminological elements from the source speech, if compared to the gold standard. This typology of errors are also treated in detail in §4.8 dedicated to Augmented Terminology. For an example of this category of errors, it is possible to examine the Non-Native speaker file 025 where, in segment 4, the domain-related term "agro-ecology" is replaced with "agriculture", as shown in Table 4.29 below.

Segment	Timestamp	Reference Speech	ASR output
1	00:00:13,600 --> 00:00:18,130	*Excellencies, distinguished guests and members of the podium*	Excellencies distinguished guests and members of the podium
2	00:00:19,310 --> 00:00:22,790	*colleagues of the UN system,*	colleagues of the UN system.
3	00:00:22,790 --> 00:00:27,930	*ladies and gentlemen, it's my pleasure to join the regional symposium*	Ladies and gentlemen. It's my pleasure to join the original symposium
4	00:00:28,040 --> 00:00:33,230	*on agro-ecology and sustainable agricultural food systems for Europe*	on agriculture and sustainable agriculture-food system for Europe

Table 4.29 – Terminology error in Non Native-speaker ASR output (extract from file 025)

Or again in the same file, in segments 24 and 26 the term "COP22" is replaced with the words/numbers "of course 22" and "22", respectively (see Table 4.30 below). All these three occurrences of Terminology errors were classified as Substitution error and were graded as Serious because they completely changed the meaning of the segment units. Given the specific, domain-related discourse, and the weight of terminology in international conferences on climate change and agriculture, terminology should in fact be considered as a relevant element to be analysed and investigated.

Segment	Timestamp	Reference Speech	ASR output
23	00:01:54,010 --> 00:01:56,930	*Today's event takes place just after*	Today's event takes place just after
24	00:01:57,440 --> 00:02:00,740	*a week of COP22,*	a week of course 22,
25	00:02:00,740 --> 00:02:03,090	*the UN climate conference in Marrakesh*	the UN climate conference in Marrakesh
26	00:02:04,130 --> 00:02:08,990	*COP22 marked an increased recognition of the importance of agriculture*	22 market and increased recognition of the importance of agriculture

Table 4.30 – Terminology error in Native-speaker ASR output (extract from file 016)

Another interesting example of Terminology error can be found in Native speaker file 051, where the technical term "Fall Armyworm" (a kind of pest for cultivations) was not recognized by the ASR system at all. Although the speaker is here Native, yet this highly technical, domain-related term was not properly recognized by the system and it was replaced with a series of phonetically similar words ("fall are more" in segment 14; "falling everywhere" in segment 16), as shown in Table 4.31 below.

Segment	Timestamp	Reference Speech	ASR output
14	00:00:49,800 --> 00:00:53,460	*Basically we want to monitor Fall Armyworm operationally,*	Basically we want to monitor fall are more operationally
15	00:00:53,940 --> 00:00:55,050	*day in and day out.*	day in and day out,
16	00:00:55,050 --> 00:00:59,160	*We'd like to know where is Fall Armyworm and we'd like to monitor its spread.*	we'd like to know where is falling everywhere and we'd like to monitor spread.
17	00:00:59,870 --> 00:01:07,480	*And it's not only FAO or countries, but it's also farmers, districts, communities, extension agents, NGOs.*	And it's not only FAO or countries but it's also farmers districts communities extension agents NGOs.

Table 4.31 – Terminology error in Native-speaker ASR output (extract from file 051)

Terminology errors are often associated with proper names of institutions, cooperation programmes, names of international initiatives or actions, specialized or domain-related terms, chemical substances, names of flora/fauna species, pharmaceutical drugs, or even with names of documents or protocols used within the international organizations. Given their domain-related or organization-related nature, these errors represent a remarkable challenge also for the automatic speech recognition of Native-speaker-held speeches. In fact, even if the Non-Native or Native speaker pronounced those terms correctly (as verified by listening to the source speech), the ASR system was not able to properly recognize them, or failed to recognize them because they are not incorporated into the built-in vocabulary (even if the system meets the LVCSR requisite seen before). Further elements of discussion will be treated in the Discussion of results and most notably in the section dedicated to Augmented Terminology (§4.8 below).

To continue with the presentation of analysis examples, it is now necessary to consider a series of errors that are quite frequent in terms of occurrences, though they do not represent errors of high severity: that is to say, the Disfluency errors. These errors generally include the misrecognition or deletion of repetitions, speech fillers, speech markers and other similar elements that are typical of discourse and orality. All these errors are largely graded as Not Serious errors in statistical terms. In file 007, for example, as shown in Table 4.32 below, the Non-Native speaker utters the "uhm" speech filler as an indication of hesitation or as his/her way of talking. Here the ASR system completely deleted this disfluent element (Deletion), without determining any loss of meaning or problem to the understanding of the segment unit in question (see Table 4.32 below).

Segment	Timestamp	Reference Speech	ASR output
1	00:00:04,030 --> 00:00:06,790	*I was always pleasant to discuss and uhm*	(I) Was always pleasant to discuss and [uhm]
2	00:00:09,600 --> 00:00:12,420	*have a conversation about many things,*	have a conversation about many things,

Table 4.32 – Disfluency error in Non Native-speaker ASR output (extract from file 007)

For another example of Disfluency error, it is possible to have a look at file 003 (see Table 4.33 below), where the Non-Native speaker repeats the preposition "to" in segment 57 as form of hesitation or as his/her way of talking. Again the ASR system automatically deleted the disfluency element of the source speech. However, though it should be accounted for an error occurrence as per the taxonomy defined for the purposes of the present study, the error had a little impact on the meaning of the segment unit (a Not Serious grading was assigned to it). In Native speaker files, these errors represent a challenge for the evaluation of accuracy by using the WER and NER models as they contribute to reducing the accuracy rate achieved when the ASR is not capable of removing them automatically. For this reason, in the present study methodology, it was proposed to adapt the NER model so as to exclude Not Serious errors from the final accuracy rate: i.e. the NER2 model.

Segment	Timestamp	Reference Speech	ASR output
55	00:04:38,710 --> 00:04:41,320	*But of course for what it is*	but of course for what it is
56	00:04:41,830 --> 00:04:46,770	*worth the FAO policy, the Voluntary Guidelines should be used*	worth the [FAO] policy. The voluntary guidelines should be used
57	00:04:47,790 --> 00:04:52,960	*to to secure indigenous peoples' rights over their lands*	to [to] secure indigenous people's rights over the lands
58	00:04:53,050 --> 00:04:56,950	*and natural resources and also of course as of any citizen to*	and natural resources and also of course as any of the citizens to
59	00:04:57,080 --> 00:05:00,020	*contribute to the national economy and development.*	contribute to the national economy and development.
55	00:04:38,710 --> 00:04:41,320	*But of course for what it is*	but of course for what it is

Table 4.33 – Disfluency error in Non Native-speaker ASR output (extract from file 003)

To conclude the survey of Fine-Grained Error examples, the Prosody category should now be considered. In speech recognition, these errors are generally associated with intonation and, according to the present study's taxonomic scheme, they account for a very few occurrences, as seen in the errors distribution shown above in Figure 4.8. More specifically, the prosodic errors surveyed in the ASR output were just represented by the deletion of question marks in speech automatic transcriptions. In file 049, as shown in Table 4.34 below, the Native speaker is giving a high-pitch, political speech and the increasing intonation includes a sequence of question marks (typical of a rhetoric style). But the ASR system automatically deleted (did not recognized) the question mark occurrences. All prosodic errors like the one shown below were not assigned a Serious grading as they did not compromised the meaning of the segments in question.

Segment	Timestamp	Reference Speech	ASR output
195	00:13:57,800 --> 00:13:59,570	*They get it right away.*	They get it right away.
196	00:13:59,570 --> 00:14:02,690	*They grasp the threat to their own future*	They grasp the threat to their own future
197	00:14:03,170 --> 00:14:07,520	*and in fact they want to be taught more about it as part*	and in fact they want to be taught more about it as part
198	00:14:07,910 --> 00:14:10,220	*of the curriculum and their normal school day.*	of the curriculum and their normal school day
199	00:14:11,390 --> 00:14:13,360	*As, are we to be content?*	as are we to content(?)
200	00:14:13,360 --> 00:14:18,110	*Are we to be content to hand down a broken planet to our*	Are we to be content to hand down a broken planet to our
201	00:14:18,220 --> 00:14:22,160	*children? That is the question we must ask ourselves.*	children(?) that is the question we must ask ourselves.

Table 4.34 – Prosody error in Native-speaker ASR output (extract from file 049)

4.4.3. Analysis of Serious/Not Serious Error category

Under this section, the distribution of the *"Serious/Not Serious"* errors (as already described in §3.6) is considered quantitatively and in relation to the study output, for the purposes of identifying specific measurements of accuracy for a live/real-time conference setting. In Non-Native speaker files, annotated *Serious* errors amounted to 935 occurrences (55%), while *Not Serious* errors were 757 (45%) in total. Unlike the previous two taxonomic levels of analysis (where the distribution of error categories was quite similar between Native and Non-Native speeches), in this case, Native speaker files are characterized by a markedly different error distribution for this categorization, with a higher number of *Not Serious* occurrences: 314 *Serious* errors (40%) and 470 *Not Serious* errors (60%). For a graphic representation of this comparative analysis, it is possible to have a look at Figure 4.9 below.

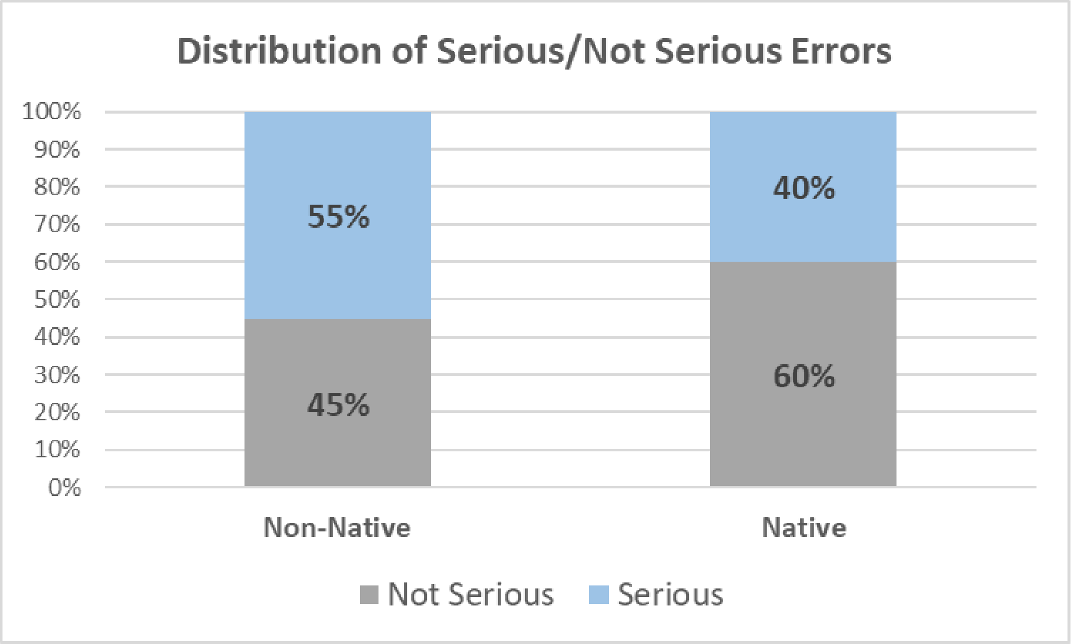


Figure 4.9 – Distribution of Serious/Not Serious Errors in Native/Non-Native files.

As mentioned by Romero-Fresco and Pöchhacker (2017) in an ASR and respeaking study similar to the present one, the degree of severity was firstly introduced in the NER model and it can be described as follows:

> "*In the NER model, the classification of errors by degree of severity is based on the extent to which a lack of correspondence between the subtitles and the original audio affects viewers' access to the original meaning, analysed in terms of (independent and dependent) idea units." (Romero-Fresco and Pöchhacker, 2017: 152)*

More specifically, under this study, the reference study's classification into 3 levels of severity (Minor, Standard and Critical) was simplified into two categories (as described in §3.7): *i.e.*, "*Serious*" and "*Not Serious*". To do so, the Critical and Standard errors as defined in the reference literature (Ibid., 2017: 153) were grouped into one single category, now simply denominated as *"Serious"*. All other Minor errors were then entered into a separate group including, by using the words by Romero-Fresco and Pöchhacker (Ibid.), those errors that *"allow viewers to follow the meaning or flow of the original text and sometimes even to reconstruct the original* *words"*. Under this category, it is therefore possible to find all errors that are not bearing significant or essential information to the segment unit or speech unit. Mostly, *Not Serious* errors are coincident with disfluency errors (errors related to the omission or substitution of disfluency elements) like the examples discussed in §4.9 below. However, they also include prosodic errors (namely, Intonation errors) and minor errors from the Grammar category (for example, the omission or substitution of articles). Certainly, the most representative example of error under this category is the presence/absence of speech fillers/markers, which for some files had a significant impact on accuracy. As already mentioned above, the solution to cope with this challenge is the possibility of measuring accuracy by implementing the NER2 rate defined in the present study: in fact, under this rate, Not Serious errors are not considered in the calculation of accuracy. The adapted NER2 rate is especially efficient in Native speaker files in which the software automatically removed those conversational elements from the final transcription output (counting as Deletion errors in NER/WER model), as also seen in previous examples above. For example, in file 016, it is possible to observe as many as 28 disfluency-based errors (over a total of 61 errors) where the speech filler *"uhm"* was omitted.

4.5. Evaluation of accuracy for VoxSigma output

In this section, the present study's analysis will focus on the evaluation of accuracy for all files automatically transcribed by using an ASR solution (namely, VoxSigma), in the attempt of identifying common features and criticalities in the ASR process. The concept of accuracy was already defined in §3.7, together with the different formulas for the calculation of accuracy rates: namely, the **WER**, **NER1** and **NER2** rates (see §3.7). The accuracy rates calculated here will be further commented in the Discussion of results (see §4.9) below, according to the different, potential applications: intralingual subtitling for people with hearing difficulties or a partial loss of hearing ("hearing impaired"), intralingual subtitling for non-hearing people (deaf) and, finally, interlingual subtitling into Italian (with the application of automatic Neural Machine Translation). Overall, this general evaluation of accuracy can also offer useful hints and evaluation considerations for the usage of ASR technology by respeakers in the production of live subtitling for non-hearing people.

In the calculations of the WER and NER rates, this study implemented a method of calculation partially adapted to the fully automatic features of the ASR system deployed, where human intervention is not included (except for the evaluation process of annotation data commented above): in fact, the role/contribution of a respeaker is not considered here. Additionally, it should be clarified that the most relevant rate for the present study is the NER rate, as it accounts for the *"Not Serious/Serious"* error severity classification described above. Furthermore, under this study, the NER rate was broken down into two different NER rates, which are renamed NER1 and NER2, for convenience, to include or exclude "Not Serious" errors from the calculation, respectively. Therefore, the accuracy NER1 rate will include the occurrences of Not Serious errors, while NER2 rate will exclude those errors totally. This should help in better representing the severity differentiation of errors and in responding more efficaciously to the various applications of live subtitling (inter-lingual and intralingual subtitling for non-hearing people and NMT application). More specifically, in the NER1 rate, *"Not Serious"* errors are assigned with a penalty of 0.5 points (*"Serious"* errors have a 1 point penalty), while in NER2 rate, *"Not Serious"* errors are not considered at all in the formula used for the calculation of accuracy (for an in-depth description of both the NER and the WER model, see also §3.7 in Chapter 3).

After these considerations, it is now possible to present the WER and NER rates so calculated for all files, in Table 4.35 below. In particular, the Table shows the *min.* and *max.* values for all three rates, including the relevant *MEAN* values and the *standard deviation*. The data refer to both the Native-speaker files and the Non Native-speaker files. It should be here recalled that the WER rate is a measure of accuracy based on the number of word error (Word Error Rate); the NER rate is based on the WER rate but it includes a categorization of error seriousness.

Values	WER	NER1	NER2
MEAN	**93.40**	**94.95**	*96.53*
MIN	**81.59**	**84.72**	*87.84*
MAX	**98.87**	**99.32**	*100.00*
STdev	**4.19**	**3.46**	*2.76*

Table 4.35 – WER, NER1 and NER2 rates for all database files.

At this point, if the database files are subdivided into two groups (Non-Native and Native speakers) as shown in Tables 4.36 and 4.37 below, it is possible to observe that Native-speaker files report a higher accuracy rate, if compared to Non Native-speaker files.

Values	WER	NER1	NER2
MEAN	**95.43**	**96.65**	*97.88*
MIN	**88.44**	**90.21**	*91.98*
MAX	**98.87**	**99.32**	*100.00*
DevSTd	**3.23**	**2.50**	*1.97*

Table 4.36 – WER, NER1 and NER2 rates for Native-speaker files.

Values	WER	NER1	NER2
MEAN	**92.31**	**94.02**	*95.79*
MIN	**81.59**	**84.72**	*87.84*
MAX	**98.20**	**98.80**	*99.40*
DevSTd	**4.27**	**3.58**	*2.87*

Table 4.37 – WER, NER1 and NER2 rates for Non Native-speaker files.

More specifically, it is possible to report that only a few files achieved a 98% accuracy with Non-Native speaker files, namely with files *002, 008, 040, 043, 045* and *053* (but with NER2 rate only) and with files *021, 023, 041* and *044* (both with NER1 and NER2 rates). The mean values for this group of files (see Table 4.37 above) are of **92.31% (WER)**, **94.02% (NER1)** and **95.79% (NER2)**, and they are all below the minimum accuracy requisite (*i.e.*, 98%). Additionally, no single file achieved the minimum industry accuracy rate of 98% with the WER rate. On the other hand, with Native speaker files (Table 4.36 above), the accuracy rate was slightly higher if compared to the previous group of files: with a **WER mean rate of 95.54%** (if compared to 92.31 WER rate in Non-Native), a **NER1 mean rate of 96.75%** (if compared to 94.02% in Non-Native) and a **NER2 mean value of 97.96%** (if compared to 95.79% in Non-Native). Yet, the minimum accuracy rate provided by the industry was not met even in the case of Native speaker files. However, it would be possible to claim that, by excluding “Not Serious” errors in the calculation of accuracy, the **NER2 average rate**

of 97.96% would be very close to the 98% threshold set by the industry and official standard of quality. Additionally, it should be highlighted that, under the Native-speakers group of files, it is possible to find a significantly higher number of single files meeting the minimum accuracy requisite with both NER1 and NER2 rates (files *020, 034, 036* and *055*) and with WER, NER1 and NER2 rates (files *047, 048, 049, 050,* and *054*). In fact, to compare these data in percentage values, the minimum accuracy requisite with NER1 and NER2 rates is achieved for 20% of total Native files (if compared to about 11% of Non-Native files) and with WER, it is achieved for 25% of the total Native files (if compared to 0% of Non-Native files).

For intralingual subtitling purposes in the source language (English), although no sufficient data are available from the present study's analysis, the files with WER and NER1 accuracy rates around 90% may however be considered as acceptable in case a respeaking process is incorporated in the workflow (not examined here), where the human intervention would allow for a simultaneous editing of subtitle units, as claimed by Romero-Fresco (2016: 59). These 90%-range accuracy transcriptions could also be considered as useful for people with a reduced hearing capacity or people with a partial hearing loss (Romero-Fresco: 2018) who are anyway capable of carrying out lip reading at a conference setting in a live situation. These transcripts would anyway represent an additional instrument for the breaking down of barriers in communications at an intralingual level.

Furthermore, for the purposes of intralingual subtitling in the source language (English) addressed to non-hearing people, as well as for the purposes of interlingual subtitling into Italian (with the application of Neural Machine Translation), only the transcription files reaching an approximate accuracy rate of 98% with NER1 rate are treated in this study (see §4.6 on Neural Machine Translation application below). In addition, it should be remarked that, with the application of Augmented Terminology resources, a strategy applied by this study (presented in §4.8), some of the transcript files from the Native group could be significantly improved in terms of accuracy, and thus be used in the interlingual subtitling process for the breaking down of communication barriers and the automatic translation into the target language, as shown at a later stage of the analysis below. Further considerations and implications from these results will be discussed in §4.9 in detail, as at this stage of the analysis, it is important to quantitatively highlight and report on statistical data only.

4.6. Analysis and comparison of transcriptions generated by GSR engine

This section presents a short analysis of the transcription data generated by Google Speech Recognition's (GSR) engine (via *YouTube* and *Descript* interfaces). The distribution of Coarse-Grained and Fine-Grained Errors (Layer 1 and Layer 2) is examined, including the distribution of "Serious"/"Not Serious" error categories. For this contrastive analysis, only a limited number of files (5 Native speaker-based and 5 Non-Native speaker-based files) is examined with respect to the Google Speech Recognition (GSR) engine. The analysis will also offer a comparative analysis of transcripts in terms of accuracy with respect to VoxSigma's output (based on the same sample files). The decision of selecting a sample of files for GSR analysis is based on a pilot test conducted initially (during the ASR technology review): the results of the pilot test showed that no significant increase of accuracy was reported with respect to VoxSigma initial test results. Objective of the present thesis is not a review of all ASR technologies available across the market.

As a first analysis of accuracy comparison between the two software solutions' output, the comparison of Perfect Match values should be considered. By referring to Figure 4.10 below, it is possible to see that Google Speech Recognition engine offered a significantly higher accuracy in terms of Perfect Match for the sample files examined (files 001, 002, 003, 004, 005, 010, 012, 013, 016, 036).

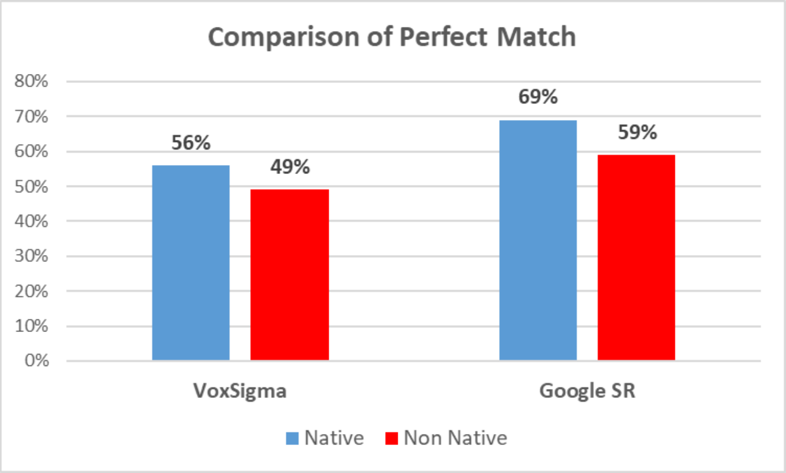


Figure 4.10 – Comparison of Perfect Match % between VoxSigma and Google Speech Recognition engine, based on the sample of files.

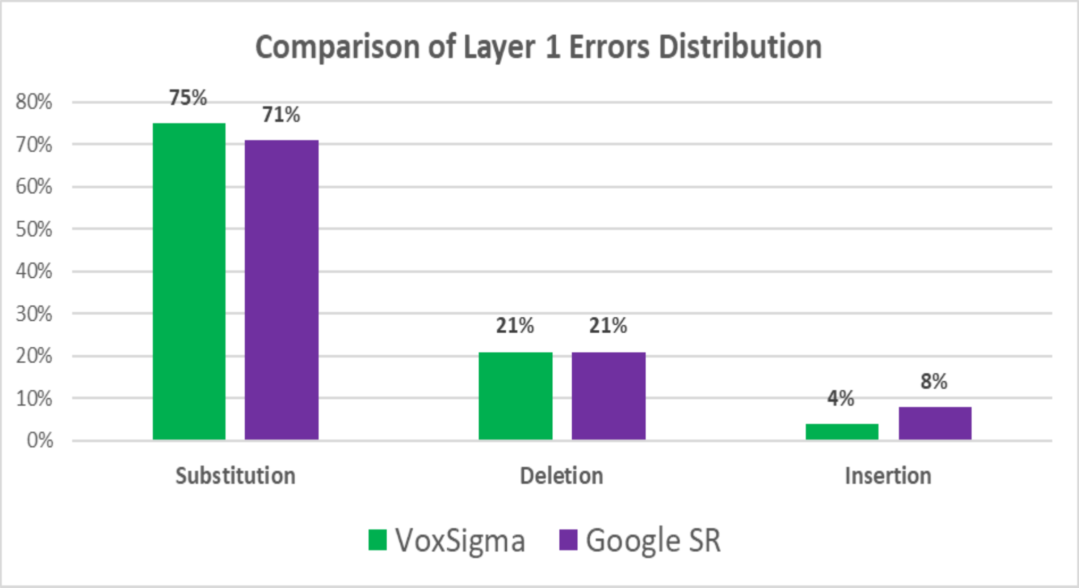


Figure 4.11 – Comparison of Coarse-Grained Errors (Layer 1) distribution for the Non-Native group, based on the sample of files.

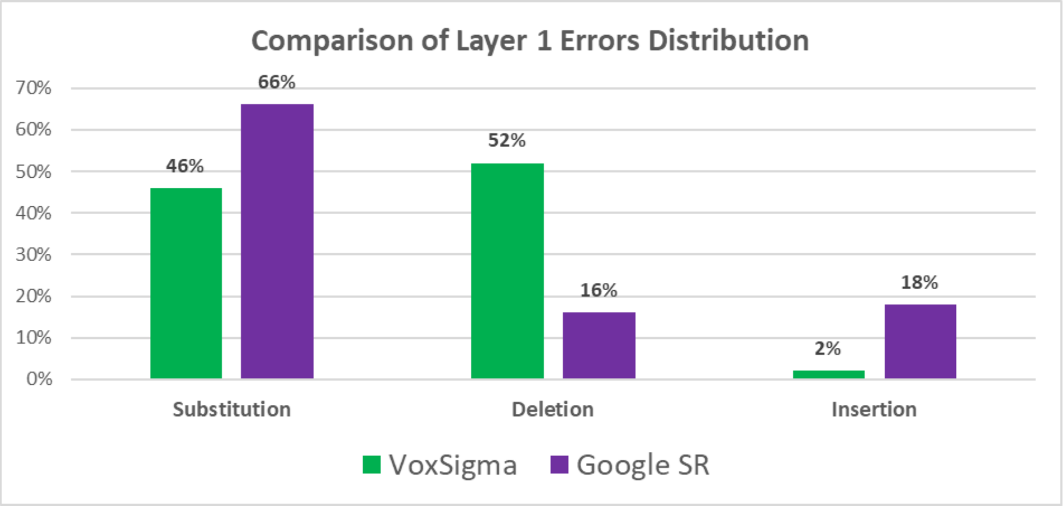


Figure 4.12 – Comparison of Coarse-Grained Errors (Layer 1) distribution for the Native group, based on the sample of files.

As far as the distribution for the Coarse-Grained Error occurrences is concerned (Layer 1 Taxonomy), as it is possible to see from Figures 4.11-4.12 above, the percentage values are roughly similar to those seen before with VoxSigma's software, with the Substitution category being the predominant one. But when considering the Native speaker files, the distribution is significantly different and it is possible to observe a higher number of Deletion errors in VoxSigma if compared to Google Speech Recognition (GSR) engine. On the contrary, GSR showed a higher number of Substitutions and Insertions. It would interesting to further examine this aspect in a more complex contrastive analysis of the two ASR systems.

Like with previous taxonomic scheme, the distribution of Fine-Grained Errors in Google Speech Recognition's output does not change in a significant way with respect to VoxSigma's one, as it is easily observable in Figure 4.13 below for the Native and Non-Native sample files, thought it should be noticed that, on an aggregate basis (provided that no significant information or evidence emerged from the subdivision above), the Grammar category covered a higher share.

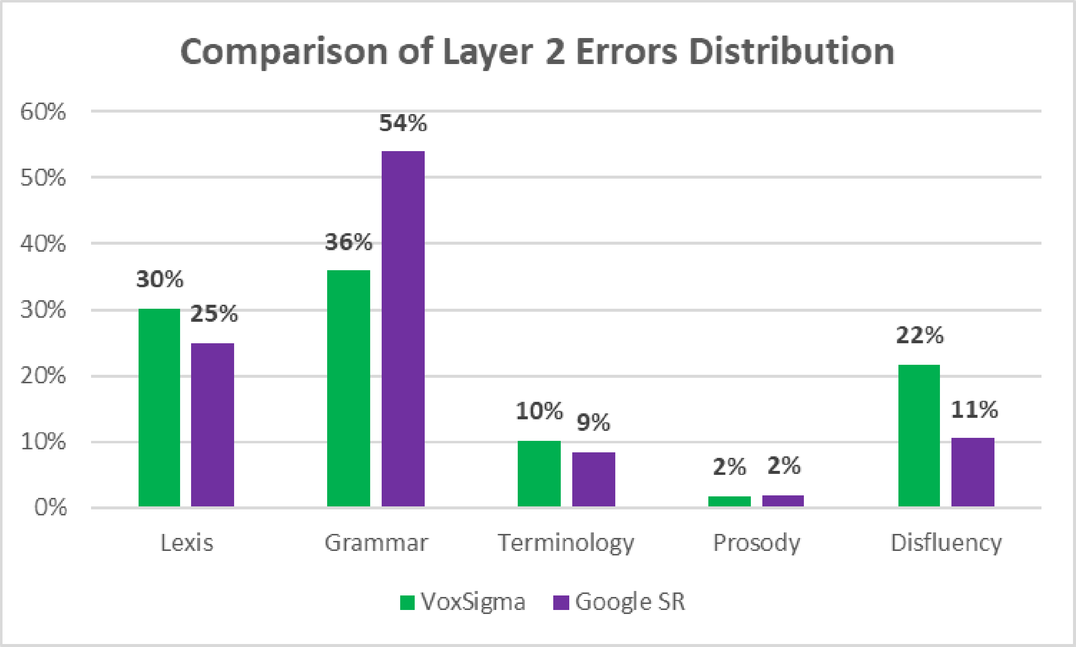


Figure 4.13 – Comparison of Fine-Grained Errors (Layer 2) distribution in Non-Native files.

Finally, in order to complete the comparative analysis, the "*Serious/Not Serious*" categorization is taken into consideration. When carrying out a further comparison with *VoxSigma*'s output, a higher percentage value for *Not Serious* errors can be observed, as shown in Figure 4.14 below.

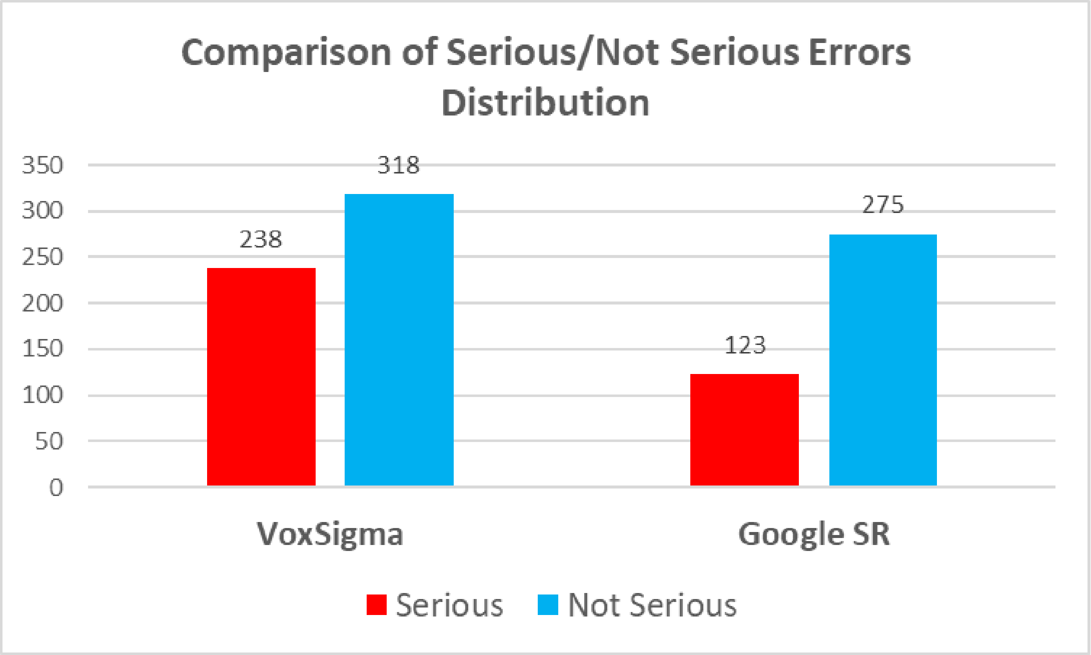


Figure 4.14 – Comparison of Serious/Not Serious Errors distribution in VoxSigma and Google Speech Recognition output, based on the sample of files.

More specifically, it is possible to comment that in Google Speech Recognition's output, "Not Serious" errors amounted to about 69% of total errors, while in VoxSigma's output, they accounted for 57% of the total. This estimate is made extracting data from the sample of files used for the purposes of this comparative analysis. Additionally, it is also possible to interpret these data as a further indication of higher accuracy when using Google Speech Recognition in the ASR process: this evaluation is also confirmed by the following analysis.

At this point, after comparing the different distributions of error categories, a comparative analysis of transcription data is performed in the attempt of examining the accuracy of GSR engine if compared to VoxSigma. In Tables 4.37-4.38 in the next page, a comparison between *VoxSigma* and Google Speech Recognition (GSR) engine's accuracy rates is provided for a sample of Native and Non-Native speaker transcription files.

File	Accuracy Rate	Engine	
		GSR	VXS
010	WER	96.81	93.73
	NER1	98.13	96.37
	NER2	99.46	99.02
012	WER	94.31	90.99
	NER1	95.73	93.36
	NER2	97.15	95.73
013	WER	95.38	95.38
	NER1	97.15	96.18
	NER2	98.93	96.98
016	WER	94.89	93.76
	NER1	97.03	96.32
	NER2	99.18	98.87
036	WER	96.96	96.96
	NER1	97.87	98.18
	NER2	98.78	99.39

Table 4.37 – Comparison of WER and NER rates for Native transcriptions

File	Accuracy Rate	Engine	
		GSR	VXS
001	WER	91.83	87.94
	NER1	94.52	90.58
	NER2	97.21	93.22
002	WER	94.44	95.29
	NER1	96.36	96.79
	NER2	98.29	98.29
003	WER	93.44	91.19
003	NER1	95.59	93.44
	NER2	97.75	95.69
004	WER	91.96	89.66
	NER1	94.4	91.77

	NER2	96.84	93.87
005	**WER**	86.17	84.04
	NER1	89.62	86.42
	NER2	93.08	90.76

Table 4.38 – Comparison of WER and NER rates for Native transcriptions

As shown in both Tables 4.37 and 4.38 above, the accuracy rates generated by Google Speech Recognition (GSR) engine (via *YouTube* and *Descript* interfaces) are slightly higher for the sample files, if compared to the VoxSigma's output, also when measuring the main accuracy rates implemented for this study. More specifically, if the mean values for the WER and NER rates obtained with both software solutions are compared, it is possible to clearly determine the accuracy rate increase, as shown in Tables 4.39-4.40-4.41 below.

WER mean value	*GSR engine*	*VoxSigma*
Non-Native	**91.56%**	89.62%
Native	**95.67%**	94.16%

Table 4.39 – Comparison of WER mean values between GSR engine and VoxSigma

NER1 mean value	*GSR engine*	*VoxSigma*
Non-Native	**94.09%**	91.8%
Native	**97.18%**	96.08%

Table 4.40 – Comparison of NER1 mean values between GSR engine and VoxSigma.

NER2 mean value	*GSR engine*	*VoxYesgma*
Non-Native	**96.63%**	94.36%
Native	**98.07%**	97.99%

Table 4.41 – Comparison of NER2 mean values between GSR engine and VoxSigma.

Approximately, the percentage increase in accuracy amounted to a span range of 1.3-1.5% for the sample of files examined. This output accuracy improvement may be of particular relevance for the selection of the appropriate software solutions in the possible configuration of an ASR system for live subtitling at public conferences or future works.

At this point of the analysis of data, during the next phase below, the application of Neural Machine Translation onto ASR transcriptions will be examined: both software solutions will be tested in the ASR+NMT pipeline for assessing the potential of Google Speech Recognition engine for the purposes of interlingual subtitling at international conferences.

4.7. Analysis of transcriptions generated by NMT

In this section of the analysis, the application of Neural Machine Translation (NMT) is evaluated in terms of accuracy and terminological coherence for interlingual subtitling (from English into Italian). The software implemented for this part of the experimental phase is *DeepL* (property of DeepL GmbH), as described in more detail in Chapter 3 (§3.9). The software solution implemented meets the advanced requirements discussed in the present study's literature review on Machine Translation technology: namely, a deep-learning neural network, cloud-based and large-vocabulary system.

As already mentioned, the NMT technology was only applied to Native-speaker files which generated a satisfactorily accuracy rate equivalent to, or above 98% under the NER accuracy rate, possibly allowing for the implementation of interlingual communications. The target language for this experimental step was the Italian language, so all transcriptions generated by ASR software were automatically translated from English into Italian via DeepL. To do so, a limited sample of transcripts automatically generated by VoxSigma and Google Speech Recognition (via *YouTube* or *Descript*) were examined: the words count for the sample files amounted to 9813 words in total and it included 6 Native-speaker files. The decision of selecting a reduced sample of files mainly depends on the fact that the marketed NMT solutions offer a reduced volume of data processing for free: also in the case of DeepL the processing of a high number of files/text would require for the payment of a fee. The analysis of accuracy was carried out by adopting a the NTR statistical model, already described in §3.9 of Chapter 3. Additionally, it should be remarked that the NTR rate also distinguishes the error severity (like the NER model) according to three categories: Minor, Major and Critical. These categories were validated by the LISA QA metric. Given the official validity of this taxonomic scheme, the present study did not include any inter-annotator agreement test for this model (it would have also represented a further effort for the annotators participating on a voluntary basis).

In order to have an insight into the nature and causes of NMT errors in interlingual communications, under the present study, it is possible to observe that most of the errors recorded in the NMT output was determined by two main phenomena: *i.e.*, **Recognition Error** (happening in the ASR part of the pipeline) and **Segmentation** (determined by the software alignment of speech text). Both concepts are clearly explained in Romero-Fresco and Pöchhacker (2017: 159) and they were validated according to the LISA QA metric. Recognition Errors (REs) are *de facto* the cause for many of the Major and Critical errors in the analysis of NMT output, and this is due to the fact that NMT cannot implement the translation process properly. And this can be simply explained by the fact that the source text/speech is not correct (parts of the transcription are wrong). On the other hand, Segmentation errors represent another important source of mistranslation, especially in relation to groups of terms or compound terms which are separated across two subtitle units. Under these circumstances, the NMT software (DeepL) could not appropriately define the context of those terms and, therefore, it could not translate them, correctly. For a better understanding of these NMT phenomena in the entire ASR+NMT system, it is worthwhile to consider some examples of both linguistic and software-related phenomena.

With reference to the phenomenon of **Recognition Error**, in file 047 for example, the speaker mentions the "*FAO*" (Food and Agricultural Organization) as the main subject of his sentence, but the ASR software (in this case, *VoxSigma*) did not recognize it appropriately in the previous step of the pipeline, as shown in Table 4.41 below, thus generating a Critical error in relation to Content in the ASR+NMT pipeline (see §3.9 for an explanation of Content errors and other typologies of errors within the NTR model).

Reference Transcription	ASR Transcription	NMT Output
we also face food security issues and	*we also face food security issues and*	ci troviamo anche di fronte a problemi di sicurezza alimentare e
nutritional issues and I think the FAO	*nutritional issues and I think if they owe*	problemi nutrizionali e penso che se devono
has been doing work in this area,	*has been doing work in this area,*	ha lavorato in questo settore,

Table 4.41 – Critical Error in VoxSigma's NMT output (extract from file 047)

When surveying file 034, another example of Recognition Error generating an error in NMT is found out as set forth in Table 4.42 below. In this specific case, the term *"FAO voluntary guidelines"* was not properly recognized by the ASR software (*VoxSigma*), and this determined a Critical error of Content in the NMT output (generated by DeepL). In the same file, also the terms *"cost catch documentation"* were not recognized appropriately in the ASR step of the pipeline, determining a Critical error in the NMT output: *"documentazione delle catture in pullman"*.

Reference Transcription	ASR Transcription	NMT Output
And we have worked with the Pacific Island countries in the development of	*And we have worked with the Pacific Island countries in the development of*	E abbiamo lavorato con i Paesi delle isole del Pacifico nello sviluppo di
FAO Voluntary Guidelines on traceability and cost catch documentation.	*PFI is voluntary guidelines on traceability and coached catch documentation.*	PFI è una linea guida volontaria sulla tracciabilità e la documentazione delle catture in pullman.
FAO Voluntary Guidelines on traceability and cost catch documentation.	*PFI is voluntary guidelines on traceability and coached catch documentation.*	PFI è una linea guida volontaria sulla tracciabilità e la documentazione delle catture in pullman.

Table 4.42 – Critical errors in VoxSigma+DeepL output (extract from file 034).

By examining this phenomenon starting from a Google Speech Recognition engine's transcription, in file 010, for example, it is possible to identify an error in NMT output due to the following Recognition Error. Here the verb *"scrumbled"* was replaced with the terms *"it's going to be"*, thus determining a Critical error for the correct translation of the subtitle unit: see Table 4.43 below.

Reference Transcription	ASR Transcription	NMT Ouput
Uhm. The World Bank's Global Partnership for the Ocean,	*Uhm. The World Bank's Global Partnership for the Ocean,*	Uhm. La partnership globale della Banca Mondiale per l'oceano,
scrumbled a little bit quite recently,	*It's going to be a little bit quiet recently*	Ultimamente ci sarà un po' di silenzio

Table 4.43 – Critical error in Google Speech Recognition+DeepL output (extract from file 010)

In the extract above, it is also possible to notice that the term "World Bank's Global Partnership for the Ocean" has been automatically translated in the wrong order of words: it seems that there exists a World Bank for the ocean. The error is not serious, but it may actually generate confusion on the target audience.

As already mentioned above, another important phenomenon of error generation in the ASR-NMT pipeline is connected with the **Segmentation** of speech parts (in other words, the subdivision into subtitle segment units), which often generates a series of errors (mostly of Major or Minor entity), but some of them with a Critical grading. This problem is due to the typical structure of speech subtitling text which is subdivided (as seen before) into subtitle segment units according to the relevant timestamp order generated in the transcription, as defined by the ASR software in the initial step of the present study's pipeline (ASR+NMT). This may certainly represent a serious challenge for the performance of a NMT engine, which considerably operates according to the context and to the words order for the selection of the target-language words, and their relevant order or distribution. An example of this phenomenon is in file 047, where the compound term *"Pacific Islands Forum Secretariat"* was distributed across 2 segment units as follows below in Table 4.44. In this case, even if the segment unit is understandable by a potential user, yet the appropriateness of terminology and its coherence is compromised. The error was rated as Minor.

Reference Transcription	ASR Transcription	NMT Output
and it had in the past a very good relationship with the Pacific	*and it had in the past a very good relationship with the Pacific*	e in passato ha avuto un ottimo rapporto con il Pacifico
Islands Forum Secretariat, which I head,	*Islands Forum Secretariat, which I head*	Segretariato del Forum delle Isole, che dirigo

Table 4.44 – Minor error in Google Speech Recognition+DeepL output (extract from file 047).

Or again in file 048, where the phrase *"to build their lives"* was perfectly recognized by the ASR software (in this case, *VoxSigma*), but it was fragmented into two segments (see Table 4.45 below). This determined for the DeepL software the impossibility of accurately recognizing the context for the word *"lives"*, thus interpreting it as the third person singular of the verb "*to live*". In this particular case, the error generated by segmentation was considered as Critical because the potential user may not understand the sense of this sentence.

Reference Transcription	ASR Transcription	NMT Output
so they have a healthy and beautiful country in which to build their	*So they have a healthy and beautiful country in which to build their*	Così hanno un paese sano e bello in cui costruire il loro
lives. Making good on the promise that each new generation	*lives making good on the promise that each new generation*	vive facendosi carico della promessa che ogni nuova generazione

Table 4.45 – Critical error in VoxSigma+DeepL output (extract from file 048)

To complete the analysis of NMT error features, it is also important to highlight that many errors can be categorized as Form error according to the model definitions, that is to say to grammar rules or style (see §3.9 on the NTR model). Even if these errors have a Minor grading in the NTR classification as they do not alter the meaning or understanding of a segment unit, yet they have an impact on accuracy as they are frequent in terms of occurrences. For an example of this, it is possible to examine the file 055 where the verb tense was not used in a coherent manner with respect to the previous segment units (see extract below in Table 4.46).

Reference Transcription	ASR Transcription	NMT Output
Now think about the shame that each of us will carry when	*Now I think about the shame that each of us will carry when*	Ora penso alla vergogna che ognuno di noi porterà quando
our children and grandchildren look back and realize that we had the means	*our children and grandchildren look back and realize that we have the means*	i nostri figli e nipoti si guardano indietro e si rendono conto che abbiamo i mezzi
of stopping this devastation, but simply lacked the political will to do so.	*of stopping this devastation but simply lacked the political will to do so.*	di fermare questa devastazione, ma semplicemente mancava la volontà politica di farlo.

Table 4.46 – Form error in VoxSigma+DeepL output (extract from file 055)

Reference Transcription	ASR Transcription	NMT Output
Yes the developed nations, that caused much of the damage to our	*Yes the developed nations, the caused much of the damage to our*	Sì, le nazioni sviluppate, hanno causato molti dei danni al nostro
climate over the last century,	*climate over the last century,*	clima nell'ultimo secolo,
still have a responsibility to lead, and that includes the United States.	*still have a responsibility to lead and that includes the United States*	hanno ancora la responsabilità di comandare e questo include gli Stati Uniti
And we will continue to do so	*and we will continue to do so*	e continueremo a farlo

Table 47 – Form error in VoxSigma+DeepL output (extract from file 050).

Or again in the same file 050, where the verb "to lead" was translated automatically in Italian as "comandare" (see Table 4.47 above), but it would have been preferable to use a term like "guidare" or "condurre". This error is classified as Minor under the study's analysis as it does not alter the general meaning or understanding of the subtitle unit, yet it gives a wrong style to the speech in a sensitive, diplomatic context like an international meeting of the United Nations.

For descriptive purposes of the NMT output, it is also interesting to analyse the categories of error according to the categories already used in the ASR process: Substitution, Deletion and Insertion. In this respect, it should be however specified that the accuracy evaluation is here made with respect to the ASR output and not with the reference transcription as it was carried in the case of ASR transcription analysis. More specifically, it is possible to see that the distribution of Substitution, Deletion and Insertion errors is similar to what was seen for the ASR transcriptions analysis, with a net prevailing number of Substitution errors (88%) with respect to the Deletion (8%) and Insertion (4%) categories, as shown in Figure 4.15 below.

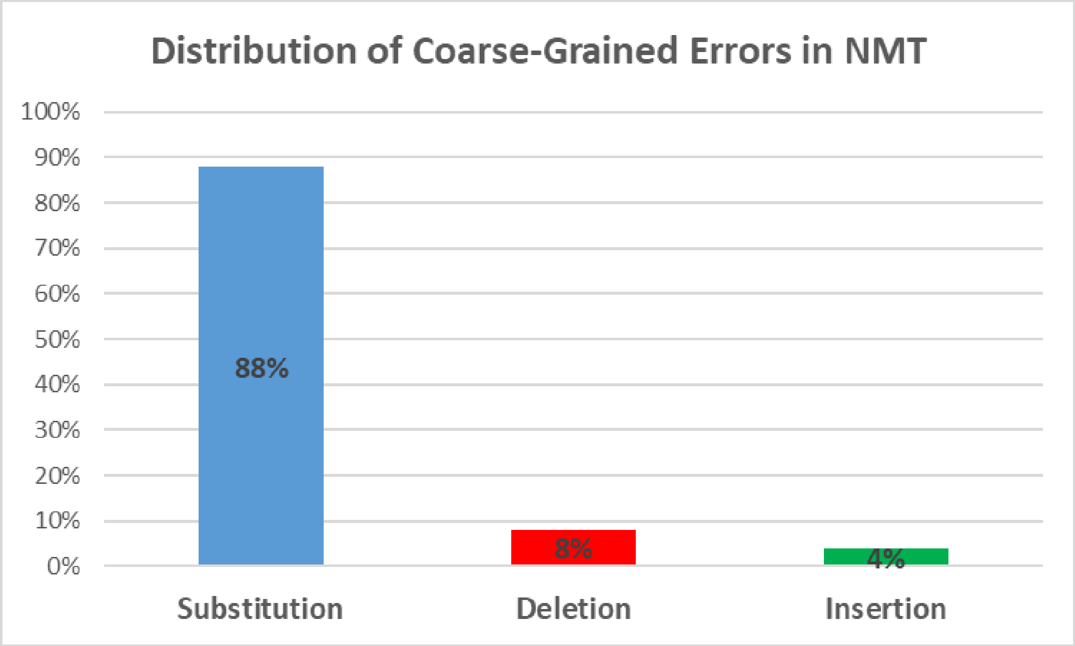


Figure 4.15 – Distribution of Coarse-Grained Errors in NMT output (with the ASR output as reference), based on the sample of files.

By examining the distribution of errors for the Fine-Grained Error categorization of the NTR model, it is possible to find out that the *Form* type errors (58%) are remarkably higher in percentage with respect to *Content* errors (42%), as shown in

Figure 4.16 below. The classification of these errors is carried out again in conformity with the NRT model described in §3.9 and approved by LISA QA metric.

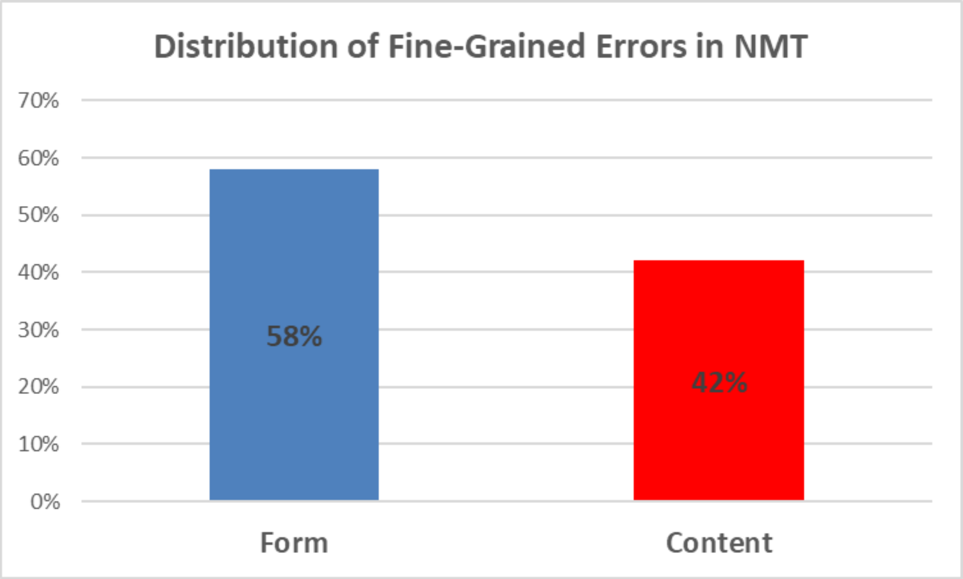


Figure 4.16 – Distribution of Fine-Grained Errors in NMT output, based on the sample of files.

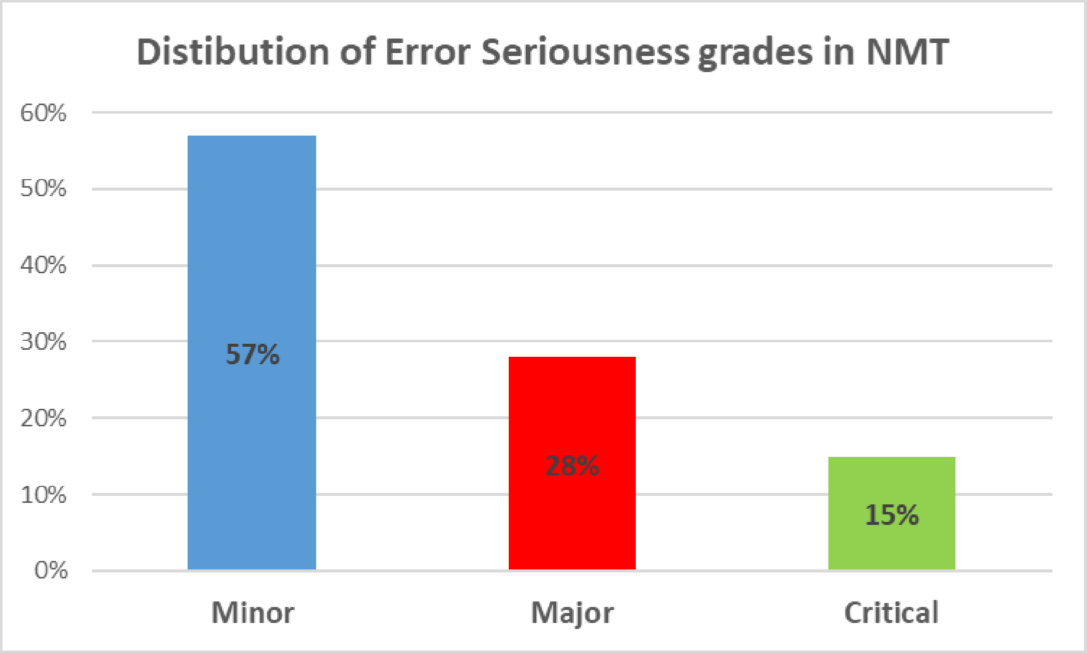


Figure 4.17 – Distribution of Error Seriousness grades in NMT output, based on the sample of files.

Finally, to complete the examination of errors as defined by the NTR model, it is possible to observe, in Figure 4.17 above, that Minor and Major errors represent the majority of errors with respect to Critical errors, which are significantly lower in percentage.

At this stage of the NMT analysis, it is now interesting to measure the accuracy of the sample of NMT-applied files by calculating the NTR rate. To do so, the study analysis calculated this rate only for a limited number of files (as listed below in Table 4.48). In particular, it is possible to claim that, with all the files examined (with Native speakers), the NTR rate was around or slightly above the 98% rate indicated in literature and required by the industry of subtitling for non-hearing people and for the purposes of multimedia accessibility, with a mean value of about 98.33%. In the Table below are the sample files with the relevant NTR rates.

File	**NTR (%)**
010	97.97
020	97.72
034	98.18
047	98.35
048	98.70
049	98.64
050	98.90
055	98.21
MEAN VALUE	**98.33%**

Table 4.48 – NTR rate for sample files

At the end of this analysis section, before drawing the conclusions from this study's analysis, it is worth considering another important element to be incorporated into an efficacious ASR+NMT pipeline, that is to say the Terminology component, as described in the next section of this chapter.

4.8. Impact of Augmented Terminology on accuracy evaluation

One of the most important novelties of this study against the reviewed scientific literature is the analysis of the role and importance of terminological resources in the processing of an efficacious ASR+NMT system and in the accuracy evaluation. As seen in previous works (e.g., in Goldwater et al., 2010), terminology-related errors in the quantitative and descriptive analysis of the final output is mainly referenced to as *"OOV – Out of Vocabulary"* errors. This feature is also reported in works by Romero-Fresco and other scholars (for example, in Romero Fresco (2016); Romero-Fresco and Pöchhacker (2017); Romero-Fresco and Martínez (2015)), where the authors only refer to this kind of issue as a decoder-related feature (this general description of the problem puts the terminology errors on the same level of other decoder-related features, like for example the latency or impossibility of recognizing peculiar feature of a language variety), without establishing a proper quantitative measure of it. To my knowledge, in all previous literature works, the so-called OOV errors are always incorporated into the macro categories of Deletion, Substitution and Insertion, without measuring statistically the real impact of this component on the final output. Hence the necessity of offering a new concept of terminology-based ASR+NMT system emerges.

As already seen in §2.2.3.2, the software solutions adopted for the purposes of this study (*VoxSigma* by Vocapia Research and Google Speech Recognition engine via *YouTube/Descript* apps) are both respondent to the Large Vocabulary Continuous Speech Recognition (LVCSR) requisite, and thus they can be considered as efficient instruments in terms of terminological coherence and representation of the specific and general vocabulary for a given language (English and Italian, for this study). But, during the analysis of data, it was evident that the decoder-incorporated terminological resources were not always sufficient to meet the automatic recognition and translation requirements of domain-specific speeches like the ones examined here. In a context-specific scenario like the international conferences on climate change and its impact on agriculture, the built-in terminological resources did in fact prove to be not sufficient. For this reason, a new concept of **Augmented Terminology** is to be introduced in ASR+NMT analysis and evaluation in order to properly cope with this challenge, which was not sufficiently examined and surveyed in literature. In my opinion, for enhanced ASR+NMT performances, the system's terminology should be augmented by incorporating a domain-specific terminology database (or more databases) which are appropriately validated and recognized by the reference bodies and institutions responsible for or organizing the institutional communications.

In §4.4.2, it was possible to learn that the impact of Terminology-related errors was of about 16% (for Non-Native speaker files) and of 17% (for Native speaker files), if compared to all other error categories, during the ASR step of the process. If it were possible to enhance the terminological resources on an *a priori* basis, it would also be possible to increase the accuracy of ASR and of NMT, accordingly. After incorporating the concept of Augmented Terminology, the pipeline for an efficient ASR+NMT system would therefore appear like the one represented below.

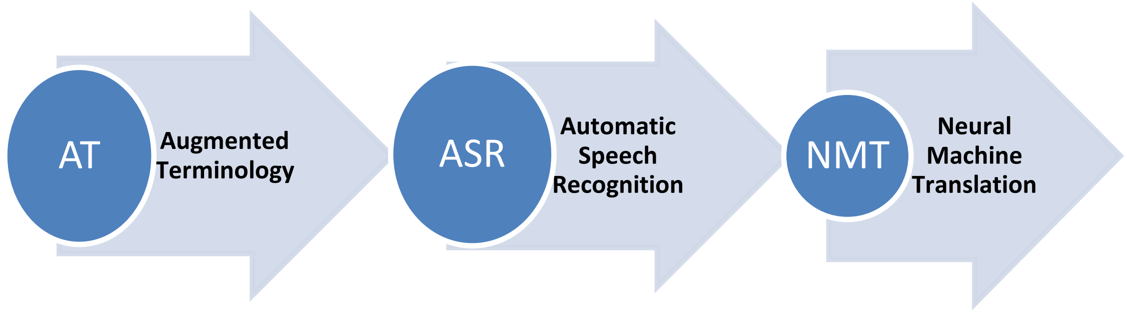


Figure 4.18 – AST system pipeline including Augmented Terminology

To better understand Figure 4.18 above, it should be added that the Augmented Terminology (AT) phase must include 1. the collection of terminology (approved and validated by the institutional body or organization) and 2. the uploading of AT database into the system. The ASR phase must include 1. the processing of automatic speech recognition via software and 2. the generation of automatic transcriptions (into the subtitle format). Finally, the NMT phase must include 1. the processing of Neural Machine Translation and 2. the reproduction of subtitles into the target language/-s.

Parallel to the definition of a new AT+ASR+NMT system, an adapted version of the statistical model implemented to measure the accuracy of output in function of terminology would be required. More specifically, this model should integrate the possibility of measuring the weight of terminology in institutional communications or media so as to identify those errors and possibly correct the ASR system deployed. This means to measure the average statistical terminological error rate for the type of conferences normally held at the institution or organization. In other words, a large-scale preparatory work would be required before implementing the system defined so far. The hypothetical model could be an adapted version of the existing NER model in which a specific measure of terminology errors could be introduced at the level of each Coarse-Grained Error categorization (Layer 1) as defined in the present study. The Terminology errors should therefore be calculated and separated from the main 3 categories: Substitution, Insertion and Deletion. However, as a definition of a new model is not the objective of the present study and provided that this operation would require further investigations and testing, here it is sufficient to mention that the AT-adapted version of the NER model would allow potential evaluators to obtain a better calculation and identification of the terminology errors weight in the estimate of accuracy. To put it simpler, the model would permit to calculate the percentage of Terminology (T) errors that could be potentially eliminated or partially corrected by applying an Augmented Terminology solution: *i.e.*, a domain-specific terminological database or vocabulary.

Given the limited, less ambitious scope of this analysis in defining a new statistical model, the present study examined the weight of terminology in two files which were selected among those having a higher percentage of Terminology errors. An experimental test was then conducted to see if those terminology-related errors could be corrected and if a better accuracy could be obtained in the ASR step of the pipeline. The analysis conducted included the use of *VoxSigma* speech recognition solution because the software permits to implement and upload an Augmented Terminology database (differently to what happens with GSR). For this step of the analysis, the terminological resources were downloaded from the Food and Agriculture Organization's FAOTERM Portal[33] and, in particular, an e-mail message was sent to the Portal's official e-mail address asking for domain specific vocabulary

[33] http://www.fao.org/faoterm/en/

databases in the area of agriculture, climate and FAO terminology. The FAO office then supplied a series of uploadable files (in particular, the IFADTERM, the Climate Change and Bioenergy database, the FAOTERM glossary and, finally, the Oceanography database) in rapid times (24 hours after the request). All these databases were delivered in the *.xlsx* format (compatible with VoxSigma platform) and they were appropriately validated by the relevant organization (*i.e.*, the FAO). After uploading the databases into the ASR solution, the analysis showed that most of the recognition errors encountered in previous processing (where Augmented Terminology was not applied) were corrected, as made clearer in the examples below.

In file 027 (a speech from Native speaker, Dan Gustafson, FAO Deputy-Director), the error problems with ASR output were mainly connected with the recognition of the term *"GAFSP"* (the official abbreviation of the term: "Global Agriculture and Food Security Program"), which is known to experts into the field and among FAO members but not included in the vocabulary incorporated into VoxSigma platform (though the platform responds to the LVCSR requisite). After uploading the IFADTERM and FAOTERM databases, *VoxSigma* was able to properly recognize that term occurrences, as shown in Table 4.49 below.

ASR without Augmented Terminology	ASR with Augmented Terminology
Excellent series. Ladies and gentlemen. Colleagues. Thank you very much for the opportunity to address the be steering committee. I regret that I am not able to be at your meeting in person. But I'm delighted that I'm able to speak by video and express how much FAO values our partnership with gas.	Excellencies. Ladies and gentlemen. Colleagues. Thank you very much for the opportunity to address the GAFSP steering committee. I regret that I am not able to be at your meeting in person. But I'm delighted that I'm able to speak by video and express how much FAO values our partnership with GAFSP.

Table 4.49 – Example of Augmented Terminology application (extract from file 027)

Across the same file, this error was repeated several times (16 occurrences) in just five minutes of speech. In addition, before the application of the Augmented Terminology (AT), the ASR system was not able to recognize other domain-specific terms such as the vey name of "*FAO*", "*IFAD*" ("International Fund for Agricultural Development") and “*SDG 2”* (“Sustainable Development Goals 2”). Thanks to the application of the Augmented Terminology solution mentioned above (i.e., the IFADTERM and FAOTERM databases), the ASR solution was now capable of recognizing those terms efficaciously, thus correcting another 4 occurrences of the terminology errors previously annotated and recorded in the analysis. This operation then permitted to obtain a higher accuracy in ASR for the file in question, raising the previous NER rate (95.60%) to 99.36% (AT-adapted NER rate), well above the minimum accuracy requisite set by the industry (and by reference literature).

By examining another sample file from this study, file 031, here the Native speaker (Allan Hruska) mentioned, in several moments of his speech, the problem of *Fall Armyworm* pest, and the ASR system was not capable of recognizing the term by means of its bult-in vocabulary. After successfully implementing the Augmented Terminology (in this case, the FAOTERM database), the ASR system (VoxSigma) could successfully cope with the recognition of this domain-specific term, correcting as many as 8 occurrences of this error, as shown in Table 4.50 below.

ASR without Augmented Terminology	ASR with Augmented Terminology
FAO has responded over the last few years, working very closely with many Member States and other stakeholders to develop a series of tools and recommendations on how to respond to (omitted) and many of them are here and the guidance notes but you can pick up or go online to the FAO fall I website which…	FAO has responded over the last few years, working very closely with many Member States and other stakeholders to develop a series of tools and recommendations on how to respond to Fall Armyworm and many of them are here in the guidance notes which you can pick up or go online to the FAO Fall Armyworm website which…

Table 4.50 – Extract from file 031: outcomes with AT application

Actually, with correcting this and other terminology-related problems (in other segment units), the accuracy rate for this file was improved in general terms obtaining a higher AT-adapted NER rate of 95.22%, if compared to the previous NER rate (90.21%) calculated before applying the AT resource. Even if the application of AT did not permit to meet the minimum accuracy rate for the industry (which is of 98%), yet it is possible to claim that a certain improvement was achieved.

In general terms, starting from this limited number of files and examples, the present study's analysis can tentatively suggest that, by applying domain-specific resources to the study's speeches, the ASR engine would be able to properly recognize the terminology available in the source audio/video files. Due to reasons of cost, the analysis dealt with only a couple of files as the processing of the entire study database would imply the payment of an extra fee in VoxSigma platform. Hypothetically, accuracy across this study database would be significantly improved, and it would be possible to meet the minimum accuracy requisite in a higher number of files, if compared to the ASR process carried out without the application of Augmented Terminology. The ASR+NMT pipeline defined can certainly benefit from this AT-based approach: in fact, several errors reported in NMT analysis derived from terminology-recognition errors (namely, Content errors in the analysis) and could have been corrected if AT resources were uploaded in the early phase. From this limited analysis of data, it was evident that the terms determining major error occurrences were those terms relating to specific vocabulary used at international organizations (for example, the abbreviations of research programmes, committee names, or initiatives) or terms belonging to specific scientific domains (for example, the names of pests, chemical substances, or specialized terminology). From the analysis conducted it is possible to observe that only the terms included in the uploaded database (specific to the organization) were corrected. Yet this final consideration would require further investigations.

4.9. Discussion of Results

In this section, a discussion of results will be presented and will focus on three main aspects: the analysis of data, and the evaluation of accuracy throughout the entire ASR (Automatic Speech Recognition) + NMT (Neural Machine Translation) pipeline and, finally, the methodology. At the end of this part, a series of claims will be enunciated with respect to the impact of terminology on accuracy evaluation. When discussing the results obtained on the analysis of data (see §4.4 and its subsections), the predominant role of Substitutions in ASR errors, the almost equivalent percentage of

Terminology errors in Native/Non-Native speakers, and the impact of segmentation on NMT should be highlighted. Further discussion should also be oriented towards the potential impact of latency and the fact that nowadays ELF is used more and more extensively at conferences.

The reason at the basis of a major occurrence of Substitution errors is probably interconnected with ASR technology itself, which tends to replace a term with another term when no match can be found in its decoder system. The difficulty in recognizing words in case of pronunciations by source speakers which differ from the ASR system's standard pronunciation can certainly contribute to increase the use of Substitutions by the ASR system. The high phonetical density (neighbourhood) of speeches and the high speech rates can also represent the cause for this phenomenon. The almost equivalent rate of Terminology errors across Native and Non-Native speakers can probably be due to the ASR decoder, which cannot recognize specific domain terms if not incorporated into the built-in language or vocabulary module. In this case, the Native/Non-Native variable has actually no effects on the overall occurrence of terminology errors. Expanding the present study with a larger database of files could be useful to investigate this phenomenon further. When considering NMT output, the problem of segmentation represents, as seen in previous section, an important obstacle in achieving the 98% accuracy threshold. Given that the present study made use of the default ASR system's segmentation, it would be interesting to carry out additional studies or testing sessions while trying to correct the segmentation. For example, the use of commas or the adjustment of wrongly truncated sentences could possibly streamline the NMT output. This would however imply the modification of the ASR system in its engineering software design, which is here used with its default configuration.

Regarding latency, the present study cannot provide data about the potential impact of latency on the target audience. However, it can be suggested that intrinsic delay in the delivery of subtitles at the end of the entire ASR+NMT process could be detrimental to the understanding and perception of the conference output by the target audience, especially if the final user makes use of other devices or strategies: for example, the lip-reading technique or signs language. In fact, there could be an asynchronous reproduction of subtitles with respect to the speaker's utterance process or the signs expressed in real time. Additionally, if the process also includes a respeaking service, this could contribute to increase the delay.

In the present discussion of results, it is interesting to examine the role of English as lingua franca (ELF) towards the achievement of accessibility. ELF is generally seen as a means to increase accessibility, but this study may demonstrate that there are important limitations. In fact, the lower accuracy achieved with Non-Native speakers suggests that EU policies (or international policies) oriented towards the expansion of ELF use do not favour the accessibility of contents in an automatic ASR+NMT pipeline like the one examined here.

To enter the discussion of results into more detail, it is possible to comment that, as already highlighted in previous studies (see, for example, Ruiz and Federico, 2014, or Goldwater et al., 2010), an "*increase in WER rate in ASR can significantly increase the so-called Translation Error Rate (TER) in the NMT output"* (Ruiz and Federico, 2014: 4). As suggested in Ruiz and Federico (2014), the analysis of data proved that *"substitutions have a greater impact (on translation quality) than deletions or insertions"* (ibid). It is interesting to observe, together with Goldwater et al. (2010), that different implications are generated across the ASR-NMT pipeline when the technology system encounters what Goldwater et al. (2010: 182) calls function words (also known as "closed class words") and content words. The former group of words is much *"more problematic for speech recognition"* according to Ruiz and Federico (2014: 10). As a matter of fact, by using the words by these scholars:

> *"The speaker may alter the pronunciation of high frequency function words, such as prepositions and articles, by under-articulating or dropping phonemes. While a human can predict these words with high accuracy, an ASR system relies on phoneme or triphone recognition as an intermediate step toward recognizing words". (Ruiz and Federico, 2014:10)*

The other problematic group of words, which Goldwater et al. (2010: 198) define as Content words (also known as "open class words" in literature), can be described, to quote Ruiz and Federico again, in their role within the ASR+AST pipeline as follows:

> *"…are generally simpler to recognize, as they often contain more syllables and cover a larger amount of speaking time within an utterance. On the other hand, open class words might not be represented in a speech lexicon, rendering them impossible to be generated by an ASR system". (Ruiz and Federico, 2014: 11)*

In this respect, the present study confirmed that Content or "open class" words proved to be more problematic, in line with the main literature in this field. In fact, as demonstrated by Vilar et al. (2006) in a similar study on ASR and SMT (Statistical Machine Translation), *"missing content words contribute more toward translation errors than missing function words"* (Ruiz and Federico, 2014: 10). Most of these errors were categorized as Lexis or Terminology errors in the ASR evaluation adopted in the present study (according to the taxonomic schemes of Layer 2), and they were often the cause of Content errors in NMT output as well, with a Critical error grading, especially when in the ASR output they determined occurrences of Substitution and Deletion errors, as also reasoned in Ruiz and Federico:

> *"Substitution errors on content words, however, have a significantly lower impact. Conversely, deletion errors on content words have a greater impact than those on function words." (Ruiz and Federico, 2014: 11)*

To recall one of the most important aims of the present study's ASR evaluation, it should be remarked that the methodology adopted aimed to evaluate accuracy and the performance of the system (as also provided in the study by Errattahi et al., 2018: 32). Furthermore, when evaluating accuracy, it should be added that, as commented in Goldwater et al. (2010: 181), speech presents features like prosody, vocabulary and disfluency factors which do increase error rates. Although it was ascertained by many scholars (for example, Lewis, 2015; Errattahi et al., 2016: 1) that ASR has significantly improved in the last years, the present study effectively contributed to evaluate accuracy (namely, the accuracy of VoxSigma and, to a minor extent, of GSR engine) so as to verify if the ASR technology may possibly be implemented at institutional levels for the breaking down of communication barriers. In particular, it is possible to comment, together with Goldwater et al. (2010: 181), that human factors or other speaker-dependent variables such as language proficiency, disfluency and canonical or non-canonical pronunciation altered the final output.

The introduction of a simple taxonomy for errors identification and annotation was indeed an important innovation of this study's methodology, if compared to the background literature, which often offered a wide array of features and error categorizations that may generate different interpretations of errors. In fact, as seen in Table 4 in §3.6, most scholars used more detailed categorizations for describing speech errors and features. On the contrary, this study attempted to produce a taxonomic scheme capable of offering neat, clearly identifiable categories of errors. In this respect, it should be highlighted that it is quite difficult to make a synthesis of all ASR criticalities and features.

For an in-depth discussion of results and a critical comparative analysis, a series of error and ASR feature examples will now be discussed, in the attempt of demonstrating the robustness of the taxonomic scheme adopted in the analysis in §4.4. and its subsections. Starting the discussion with the Substitution category, it is possible to comment that these errors may be due to four main reasons: the speed rate of the speaker (preventing the decoder to correctly recognize the exact words), the speaker's pronunciation with respect to the correct English pronunciation (as specified in Chapter 2, the correct English pronunciation is the English variety incorporated into the ASR system: i.e., the UK or US English varieties), the density of the text, the phonological neighbourhood, and the absence of that term in the software acoustic and language model. In our transcription output, this phenomenon occurred more frequently when the software could not identify and recognize the proper names of individuals or the proper names of institutions, organizations, programmes, initiatives, etc. For example, the program name *"FAMEWS"* of the FAO (Food and Agriculture Organization) was replaced with the adjective *"famous"* or the pest name term *"Fall Armyworm"* was replaced with the terms *"fall I"*.

Under the Substitution category of this study, Goldwater et al. (2010: 195) identify another specific ASR feature: "*many of the other high-error words involve morphological substitutions"*. This kind of phenomenon mainly regards the bare stem and the grammatical declension or conjugation of verbs (and also of adjectives: e.g. *"high/higher"*) and it is intrinsically connected with the acoustic and language model, which is *"often insufficient to distinguish these two forms* (for example "high/higher") *since they can occur with similar neighbouring words"* (Goldwater et al., 2010: 195-196). A source speech example of this kind of error is "*call/called"*, "*asks/asked"*, "*happen/happened",* etc. For this kind of errors, in the study's ASR transcription output, it was possible to find out a plethora of examples (all under the Substitution occurrences as per the taxonomy Layer 1) and, for this reason, it is possible to claim that, together with other substitution-related phenomena (described in literature as Phonetic Substitution and the Homophone/Near-Homophone features), it represented a high share of the Substitution error occurrences. In file 038, segment 48, it was possible to find, for example, the verb *"accomplished"* replaced with its verb bare stem form, *"accomplish"*.

Within the Substitution category, homophone represents another challenge for the ASR system (see, for example, Romero-Fresco and Pöchhacker, 2017; Goldwater et al., 2010). This subcategory of substitution error is easily explainable, and it happens, in a few occurrences, also across this study's transcription output. Generally speaking, these errors mainly concern with words or terms having identical phonetic sounds like, for example, in the words *"seat/seed"* or *"dessert/desert"*. Though being a low-frequency phenomenon in the study's output, this typology of wrong recognition had a certain impact and it was mainly due to the fact that the context for the software to be automatically processed was often limited by the segmentation of text offering reduced context information to the ASR system; for this reason, the ASR could not recognize and easily disambiguate between the two terms. This issue was also commented in the previous chapter of the analysis (§4.8). Examples of this error were found in file 022, where the verb *"see"* was replaced with *"sea"* or, in file 016, the term *"roots"* was replaced with *"routes"*. Similarly, near-homophones and neighbours represent example of Substitution errors (see Goldwater et al., 2010; Mirzaei et al., 2018; Luce and Pisoni, 1998; Vitevitch and Luce, 1999). This highly-frequent occurrences of error appeared in the ASR output when a near-homophone term replaced a given term which had a near phonetic sound. Generally, as highlighted by Goldwater et al.:

> *"The context in which a word is spoken is sufficient to disambiguate between acoustically similar candidates, so competition from phonetically neighbouring words is not usually a problem." (Goldwater et al., 2010: 195-196)*

But when we have *"doubly confusable pairs"* (Goldwater et al., 2010), *i.e.* words with one or two strong competitors that may be used in similar contexts, the Substitution error becomes more frequent. In fact, as underlined in Goldwater et al. (2010: 196): *"word pairs with similar language model scores in addition to similar acoustic scores can be a source of errors"*. Examples of this typology of error were, to mention just a few of them, the pairs: *"than (and)"*, *"then (and)"*, *"him (them)"*, and *"them (him)"*. Across this study's ASR output, it was possible to find Substitution errors like *"face/faith", "these/this", "we/with", "won't/want"*, etc.

The Substitution error category can also be associated to another typical behaviour of ASR software, which is denominated "Lexical Frequency" (see for example the works by Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001; Gada et al., 2013). To put it simply, this feature can be described as the implementation, across the transcription output, of the most probable or most frequent term or expression in a given context (for that language) when the software cannot properly recognize that term for various reasons (for example, the high speech speed of the speaker, a wrong pronunciation of the term or the density of the text). This kind of behaviour may thus generate an error in the form of a Substitution with the usage of a more common or more frequent (and then more statistically probable) term. To say it with Goldwater et al. (2010: 182), *"ASR systems are better (faster and more accurate) at recognizing frequent words than infrequent words"* and this is also true for human speech recognition. Again, with Goldwater et al. (2010: 190), it is possible to highlight that: *"we find that low-probability words have dramatically higher error rates than high-probability words"*. In the transcription output, this phenomenon was found out, for example, in file 011 where the term *"fingerlings"* was replaced by a more frequent term, *"finger"* or in file 018 where the term *"afforestation"* was replaced with the terms *"Air Force station"*.

Under the Substitution category (but to a minor extent also in the Deletion category), another criticality of the ASR system is represented by the Terminology errors (reported also in Romero-Fresco and Pöchhacker, 2017; in Gada et al., 2013). This phenomenon is often associated in literature to the misrecognition of a common noun or specialized term without examining the reason for it. Under this phenomenon, the software deletes or replaces a term which is not available in the software language model and vocabulary with a more probable or frequent term. To the best of my knowledge, after examining results from the present study, the real problem with this type of error stands in the level of specialization of the term in question. For example, in file 018, the term *"AFOLU"* (abbreviation for the "Agriculture, Forestry and Other Land Use" initiative by FAO) was replaced with *"a for-LU"*; the abbreviation *"SIDS"* (standing for Small Island Developing States) was replaced with the term *"seeds"*; or in file 022, the term *"SDGs"* (indicating the Sustainable Developing Goals) was replaced with "*GS"*. In the present study, these errors were quantitatively accounted for in Layer 2 Taxonomy under the Terminology, according to the criteria set forth above.

To continue with the discussion, under the Substitution/Deletion category, another source of criticality that affects ASR systems is represented by the identification and recognition of numbers/dates (Romero-Fresco and Pöchhacker, 2017)**.** In a conference environment with the presentation of scientific data and discourse argumentation based on numerical values (like the present study's climate change and agriculture focus), this typology of error posed a serious threat to the correct recognition and transfer of information to the audience. In file 039, for example, the year date *"2050"* was replaced with *"2015"* or, again, in segment 58, the number *"8"* in *"eight scenarios"* was deleted; in file 022, the percentage value *"80%"* was substituted by *"18%"*.

Under the Deletion category, a feature which is worth being discussed is probably intonation. In the present study, the intonation feature is coincident with the omission of the question mark (*"?"*) in the transcription output. This feature was mainly due to the segmentation of text units, the impossibility for the software to recognize intonation or, often, to the incorrect intonation pronounced by the speaker (especially in the Non-Native speakers group of files). As already commented above, the correctness of intonation is based on the ASR system's standard English varieties. Prosody errors were mostly conditioned by the speaker's way of talking, as confirmed in Goldwater et al. (2010: 196). Though with a few occurrences, this ASR error can be found in files 001 (segments 11 and 76), 005 (segment 22), 010 (segment 39), 016 (segment 84), 036 (segment 14), 042 (segment 20), 049 (segments 199 and 201) and 051 (segment 13).

To continue with the discussion of results and the main features and criticalities concerning coarse-grained errors, under the Deletion category (and also, marginally, under the Substitution and Insertion categories), it was possible to detect several disfluency errors in the study output. In particular, these features includes speech features such as "Start/End of speech", "Speech Fillers/Markers", "Hesitation/False Starts", and "Repetitions" (Goldwater et al. (2010); Ruiz et al. (2017); Adda-Decker and Lamel (2005); Gada et al. (2013)). Most of the errors related to these speech features were considered as Not Serious errors. Though these features of the ASR process had actually a low impact on the accuracy levels of the transcriptions, yet statistically they accounted for significant high percentage values over the total number of errors. In particular, to quote Goldwater et al. (2010: 198), it is possible to claim that *"disfluencies heavily impact error rates"*, if accounted for in the calculations. An example of the first feature above (Start/End of Speech) can be found in file 051 where the phrase *"let me"* was deleted by the ASR software at the beginning of the speech, or in segment 186 of the same file where the final thanks giving (*"thank you very much"*) were omitted at the end of speech. The Speech Fillers/Markers (also called "discourse markers") feature had strong effects on the evaluation of accuracy as these elements are a typical feature of orality and, statistically, they accounted for a significant percentage of errors in this study (as seen in the analysis of Disfluency category, §4.4.2). Yet, provided that they did not bring essential or significant information into a speech unit, in most cases they were considered as *"Not Serious"* errors, and as such, having a minor impact on the accuracy rates. An example of Disfluency errors, it is possible to consider file 001, at segments 40, 44, and 78 where the speech filler *"uhm"* was deleted by the software. As far as the Hesitation/False Start feature is concerned, it should be observed that occurrences for this feature were statistically lower in number, and they had a minor impact on the total number of Disfluency errors. Generally, these errors occurred when the speaker was uncertain about the formulation of his/her information or sentence and when the speed rate was low. But trying to identify common characteristics for this kind of error is not possible as hesitation is strictly interconnected with the speaker's way of talking and no generalizations can be obtained from the study output. For an example of this feature, it is possible to review file 017 where conjunction *"and"* was omitted: here the speaker was hesitating in the formulation of his speech, and he changed the discourse. It should also be commented that this feature was often strictly interconnected with the usage of speech fillers like, for example, *"well"*, *"uhm"*, etc. So distinguishing them from other disfluency-based errors is not an easy, unambiguous task. For this reason, for the purposes of statistically quantifying them, all these errors were grouped together under the Disfluency category. Finally, to conclude the discussion of the most common features of coarse-grained errors, within the disfluency-based subgroup, it is possible to examine the Repetitions feature. This element, which is typical of orality, is generally used by a speaker to emphasize information or a concept, or, alternatively, when there is a certain hesitation or confusion in the sentence formulation: reformulating or adding new pieces of information is a typical behaviour in a speaker's way of talking. As in the case of hesitation and of speech fillers or markers, repetitions were often omitted by the ASR software, contributing to the statistic quantity of Deletion category errors. For an example of that, it is possible to consider file 021, at segment 21, where the speaker pronounced the words *"…have made to the market, the international market"* but the ASR software deleted the first occurrence of the term *"market"*; or again in file 016 (at segment 3), where the repetition of the pronoun *"I"* in the sentence "*Sorry, I, I don't think it would be wise for me to follow…"* showed a certain hesitation by the speaker and the software automatically eliminated it.

At this point, after having examined practical examples of ASR errors and having offered a detailed discussion of the results, it is important to verify if the Research Questions expressed in Chapter 3 of this study were responded or partially responded. To do so, the initial **Research Questions (RQs)** are now recalled below:

1. Can ASR technology produce accurate output for the breaking down of the barriers of communication in the intralingual context (in the English language)?
2. Can the combination of ASR and NMT provide an accurate output in generating subtitles for the purposes of accessibility in the interlingual context (namely, from English into Italian)?
3. Do domain-specific terminological resources (incorporated into the ASR step of the pipeline) improve the accuracy of interlingual and intralingual subtitles in this study's specific scenario?

With regard to the **first RQ**, it is possible to maintain that, on the basis of the results obtained in the analysis phase, the examined ASR technology proved to be partially successful in achieving an accurate output for the database of files included in this study. More specifically, it is possible to comment that, with Non-Native speaker files, both VoxSigma solution (property of Vocapia Research) and Google Speech Recognition engine (via YouTube or Descript interfaces) were not able to reach the predefined, industry-standard minimum accuracy rate of 98%, though they both obtained a substantially high rate of accuracy under the two reference models of the evaluation models adopted. In fact, in the case of VoxSigma-generated transcriptions from the Non-Native speaker files, the accuracy was as follows: 92.31% with WER, 94.02% with NER1 and 95.79% with NER2. In the case of GSR engine transcriptions for the sample examined (based on 10 files sample) was: 91.56% with WER, 94.09% with NER1, and 96.63% with NER2. On the other hand, with the Native-speaker files, both solutions obtained significantly higher accuracy rates, almost approaching (and achieving in the case of GSR) the 98% threshold set by the industry standard and by reference literature with the NER2 rate. In fact, with VoxSigma-generated transcription, the accuracy was of: 95.43% (WER), 96.65% (NER1) and 97.88% (NER2). In the case of the sample of files examined for GSR-generated transcriptions, the accuracy (based on the 10 files sample) was of: 95.67% (WER), 97.18% (NER1) and 98.07% (NER2). For intralingual communication purposes, it should therefore be commented that, with both groups of speakers (Native and Non-Native) under this study, the ASR technology, though responding to all the technological requisites seen in Chapter 2 (§2.2) and Chapter 3 (§3.4.1), actually failed to effectively meet the minimum accuracy rate. Yet, by taking into consideration the fact that the accuracy rate was mostly determined by Not Serious errors in the case of Native speakers, as seen in §4.4.3 of the analysis (Chapter 4), it is possible to claim that, for this group of speakers only, both software solutions succeeded in meeting the industry's predefined threshold for accuracy, with a mean value of 97.88% in the case of VoxSigma and 98.07% in the case of GSR (under the NER2 model). In fact, given the minor weight of not-serious errors in the understanding and meaning of the segment units and the entire subtitles contents for those speeches, it is absolutely plausible to consider those subtitles to be sufficiently understandable by a potential user or viewer.

In addition to the discussion above, considered the possibility for many of the targeted audience (non-hearing people or persons with partial hearing loss) to use the so-called "lip reading" technique, or even the possibility of potentially introducing a respeaker into the process of speech recognition and editing (not examined under this study), it is possible to believe that the output generated by ASR technology can certainly prove to be a valuable, additional instrument for the breaking down of the communication barriers across the targeted users/viewers. In fact, by means of lip reading and/or the intervention of a respeaker during the speech recognition phase of the pipeline examined here, it would certainly be possible to obtain higher levels of understanding by part of the final users and higher accuracy rates, also in the case of Non-Native speakers. In this respect, it should be commented that the studies on interlingual respeaking (especially in the field of Media Accessibility) pose a series of challenges and share many aspects and issues with the present study, where speech recognition is involved in the generation of subtitling. This study should therefore refer to, and possibly contribute to expand, in future works, the approach adopted in the ILSA project described in Chapter 2 (§2.4), which had probably the merit of identifying the role and impact of a respeaker and live subtitler in the ASR process and responded to the needs of a wider audience of physically-impaired users. In many studies, the deaf minority is to some extent left behind to the benefit of a majority of hard-of-hearing viewers. It is therefore of utmost importance to produce accurate subtitles and *"ensure that wider access does not involve lower quality"*, as highlighted by Romero-Fresco (2018: 192). Finally, with respect to the selection of an effective ASR technology for the process, it is possible to add that, as described in §4.6, Google Speech Recognition engine proved to offer better outcomes in terms of accuracy, if compared to VoxSigma. Yet it should be made clear that the present study was not aimed at reviewing all marketed ASR technologies and that other ASR technologies may probably offer better results or performances.

As far as the **second Research Question** is concerned, it should be highlighted that the discussion below only refer to the communication scenario examined here (international conferences on climate change held by single speakers) and that the target language is Italian. Additionally, a distinction between Native and Non-Native speaker-held speeches should be made again like for RQ 1. In fact, the accuracy rate obtained in this study pipeline (Automatic Speech Recognition + Neural Machine Translation) was only calculated starting from speech files that previously met the minimum accuracy rate under the ASR phase and for a limited sample of files including Native speakers only. Thus the ASR + NMT pipeline defined here was tested only for the Native group of files and the entire system was not deemed, on an *a priori* basis, to be successful for the Non-Native speeches: in fact, the application of NMT to files where the minimum accuracy rate was below or far below 98% could only make the final output even worser in terms of accuracy, as further NMT errors would be expected to be added to those generated by ASR technology. By examining the Native sample of files presented in §4.7 above, it is possible to highlight that the NTR rate achieved was 98.33% for the sample of files examined. With these results, it is therefore possible to claim that with Native speaker-held speeches, the ASR technologies deployed here offered the possibility of completing the entire process of communication and translation into the target language (Italian), contributing to further break down the barriers of communication for non-hearing people (but also for other potential users/viewers) for effective interlingual subtitles and communication.

To respond to **RQ3** above, it is possible to claim that this study examined the ASR output in an innovative way with respect to previous, reference literature studies where domain-related or technical terminology was regarded only as "out of vocabulary" elements of a given speech (see discussion above). As seen in previous works (e.g., in Goldwater et al., 2010), terminology-related errors in the quantitative and descriptive analyses of the final output are mainly referenced to as *"OOV – Out of Vocabulary"* errors. A mentioning of this feature was also reported in other studies by Romero-Fresco and other scholars (for example, in Romero Fresco (2016); Romero-Fresco and Pöchhacker (2017); Romero-Fresco and Martínez (2015)), where the authors only referred to this kind of issue as a decoder-related feature, without establishing a proper quantitative measure of it. To my knowledge, in all previous literature works, the so-called OOV errors were always incorporated into the macro categories of Deletion, Substitution and Insertion, without measuring statistically the real impact of this component on the final output. Hence the necessity of offering a new concept of terminology-based ASR+NMT system emerges. De facto, this study analysis did indeed examine the impact of terminology both at the level of Fine-Grained Error categorization (Layer 2 Taxonomy, analysed in §4.4 and its subsections) and, most notably, as a specific Augmented Terminology (AT) component to be integrated into an adapted version of the NER model (see §4.8). More specifically, the study offered the possibility of surveying the impact of terminology-related errors throughout the entire database by highlighting the necessity of introducing domain-related terminological resources into a AT + ASR + NMT system so defined. In particular, it was evident that accuracy rate could significantly rise, for the two examples examined, with an increase by about 4.50-5.00% in an AT-adapted NER rate. To sum up, it is possible to claim that terminology resources can improve the final output accuracy in the setting and communication scenario described in this study, both for Native-speaker held speeches and for Non-Native-speaker held speeches. More specifically, advance preparation is considered one of the most important activities to ensure quality in the usage of ASR, especially in the case of highly specialized domains: this consideration finds a certain grounds in the works by Kalina (2005) and Gile (2009), who described the necessity of a preparatory material activity for the interpreters. As commented by Xu (2015), the use of precise terminology can in fact enhance the communications in interpreting services, but, to my judgment, this can also be applied to an ASR + NMT system process.

After having discussed in the paragraphs above if the RQs were responded, a series of considerations should be made with respect to the methodology, analysis and evaluation of results, in order to further substantiate the results claimed above and to highlight the potential weaknesses and strengths of this study.

By referring to the definition offered by McCowan et al. (2005: 2), it is possible to comment that an ideal **ASR evaluation methodology** should be *"direct, objective, interpretable and modular"*. From the methodology described in more detail under Chapter 3, it is evident that this study (as also developed further in Chapter 4) responded substantially to all these four criteria. In fact, the ASR evaluation methodology adopted here can be considered as being direct because the measurement of the ASR output was carried out independently of the ASR application used: that is to say, the results were not examined according to or by means of the ASR technology itself, nor were they based on the relevant ASR technology selected for the processing in that given moment. Criticism of the present study might negatively highlight the limited number of ASR technologies implemented, but, as explained in Chapter 3, other two solutions were reviewed on a preliminary basis: however, they were rejected for not responding to the minimum ASR requirements of the sector (i.e., Dragon Naturally Speaking by Nuance) or for the high-cost associated to its usage (Microsoft Skype Translator). Regarding the limited implementation of GSR engine in this study (the solution was used for automatic transcription of 10 files only), it should be clarified that the objective of this study was not to identify the best-performing ASR solution, nor to review all the marketed ASR products. Additionally, the ASR evaluation methodology adopted in the study proved to be objective as the value of accuracy was calculated in an automated manner. In fact, the data from the annotation process were calculated and quantified by using the validated WER and NER model rates. The methodology also responded to the interpretable requisite as the accuracy rate so calculated (the measure) was also an indication of the performance offered by the ASR technology examined (notably, VoxSigma). In this respect, other potential instruments or measures might have been used in the analysis (as better described below) but the selection of the WER and NER models (for the ASR evaluation) and the NTR model (for the NMT evaluation) proved to offer easily interpretable and objective information on accuracy. Finally, the methodology of the present study can also be considered as being modular, as the analysis offered both general accuracy data (WER, NER and NTR rates) but also other sub-measures to be calculated starting from the general basic data: for example, the NER2 rate or the AT-adapted NER rate were a result derived or based on the general NER accuracy rate. Critiques to the present study may be moved with respect to the parameters used in the evaluation of accuracy. Other metrics could probably have been used. With reference to the BLEU metric, it should be commented that, notwithstanding its consideration as a benchmark standard for automatic evaluation of MT output, it is also accepted (Way, 2018: 168; Koehn, 2009: 229) that it actually presents a series of limitations. More specifically, with the BLEU metric the source text and the reference translation are ignored in its calculation (Way, 2018). On the contrary, the WER, NER and NTR rates used in this study are based on reference transcription or translation (the so-called "Gold Standard"). In fact, although the automatic evaluation methods are often considered as more accurate and objective because they limit the usage of human intervention (Castilho et al., 2018b), it is important to state that automatic evaluation methods strictly require translations (in the case of NMT) or transcriptions (in the case of ASR) carried out by humans or professionals, the quality of which is not verified (Castilho et al., 2018b). The so-called Gold Standard is considered to be correct on an *a priori* basis under the present study. However, the risk of errors in the manual transcription and human imprecision are high and were not probably examined on an enough substantial basis. In this respect, it should be however specified that the manual transcriptions of speeches were counterchecked by a mother-tongue interpreter when the audio or text of the source speech was not clear. Together with Castilho et al (2018b), it should also be remarked that, in this study's evaluation methodology, the analysis and annotation process is carried out at the level of single segment units (as by the default ASR segmentation) and this may imply a minor precision in the evaluation of output coherence in terms of terminology.

Furthermore, as already seen in Chapter 2, to complete the discussion on the methodology, the background literature also offers two important requisites to be met for an evaluation methodology to be effective: i.e., to be "rigorous" (research-informed, valid, reliable, user-focused) and "transferable" (straightforward, flexible and valid for training), as proposed in Romero-Fresco (2020). To start with this point of discussion, during the annotation process, a high degree of subjectivity may have intervened in the evaluation of ASR output. The problem of subjectivity is often at the heart of the debate on quality assessment within the translation studies and, parallelly, within Media Accessibility and the subtitling industry. And this issue can be better coped if the methodology adopted in the evaluation responds to the two criteria mentioned above. In particular, this study's methodology and accuracy evaluation models can be considered as sufficiently rigorous for being, first of all, research-informed (i.e., based on previous research). In fact, when considering one of the most widespread models of quality assessment in subtitling for Media Accessibility, the NER model, it is possible to assert that its formula is derived and mostly based on the basic principles of the WER (word error rate) model, as officially approved and validated by the US National Institute of Standards and Technology and on its adaptation by the Centre de Recherche Informatique de Montréal (CRIM) (see Pablo Romero-Fresco, 2016). In the same way, the NER1, NER2 and AT-adapted NER rates presented in this study can be evaluated as research-based as they are effectively based on previous, approved models. Also with respect to the classification of errors in terms of severity (Serious or Not Serious), it is possible to underline that the categorization is based on previous works and, most notably, on the research project set up in 2010 by the Carl and Ruth Shapiro Family National Center for Accessible Media (Apone et al., 2010) and especially on the findings of the EU-funded DTV4ALL project (Romero-Fresco, 2015). Secondly, the methodology deployed in the study is rigorous for being recognizable as a valid model of ASR evaluation. By taking into consideration, for example, the WER rate, the parameters and dimensions which are measured (i.e., accuracy, speed rate, Native/Not-Native fluency in English), are agreed on the basis of official consultations by governmental regulators in the UK and Australia with broadcasters, subtitling companies, researchers and user associations (as reported in Ofcom, 2015) or they do represent parameters with a predetermined definition (as in the case of the speed rate, which was calculated according to the industry's wpm rate). Yet in the assessment of accuracy, a certain degree of human intervention was required to verify, for example, if a loss of information was to be accounted for in the evaluation of the final results. Additionally, to mitigate the degree of subjectivity introduced by such human intervention, the inter-annotator agreement test was set up, which further substantiated the validity of the taxonomic scheme. By means of this instrument, it was in fact possible to offer further grounds to the taxonomy scheme adopted and to the methodology implemented, in addition to responding to the requisite of reliability expressed in literature, as commented in the next paragraph.

A key element for the **reliability** of a model of ASR evaluation is certainly the calculation of the inter-rater or inter-annotator agreement rate (or IAA rate) between different evaluators. The test conducted within the Department of Interpretation and Translation of the University of Bologna (for the purposes of this study) was based on previous, similar tests, like for example the Live Respeaking International Certification Standard (LiRICS) initiative, and being also a research-informed test, this contributed to consolidate the reliability of the results obtained. Criticism against the present study may point to the fact that the pool of annotators selected (7 annotators plus the main annotator) was not sufficiently varied, it did not include experts in ASR technology and it was involved in the annotation of 2 audio/video files only. Certainly, this aspect may represent a challenge for the test validation, but probably, in my opinion, the fact that researchers with no or limited expertise in the field of ASR technology were recruited may actually add further solidity to the results, as it may be tentatively suggested that higher IAA rates could have been achieved if annotators were trained or qualified experts in the field (as in the LiRICS initiative). After all, the IAA rates obtained were substantially high, given also the not-so-expert pool of annotators involved.

As already mentioned in another part of this work, a **rigorous methodology** for the quality assessment in ASR, but also in Media Accessibility and the subtitling industry, is expected to be user-focused, in line with what Greco calls the second of the three shifts produced by the accessibility revolution: *"the change from a maker-centred to a user-centred approach"* (Greco, 2018). In the present study, the requisite was not met when considering the exact role played by the final users/viewers of subtitles, though the consideration of the accessibility was at the centre of the analysis. In fact, different degrees of error severity (and thus the final score) assigned to each error may be considered as an attempt of evaluating accuracy for the final understanding of the targeted audience. The seriousness score was in fact based and assigned exclusively on the factual understanding of the single segment unit or of the entire subtitle contents by the present study's author. However, the weakness is certainly represented by the fact that the targeted users were not involved in the evaluation of the final output. This could have been produced by generating user-oriented questionnaires on the quality or accuracy of subtitles. De facto, the on-screen visualization and the latency of subtitles could hamper or make more difficult the understanding of the segment by the target users.

Regarding the rigorousness of the ASR evaluation methodology applied, critiques may be moved against the present study on how or if this methodology is sufficiently solid to obtain an impact on society, that is to say if it can have a certain utility. In fact, according to Romero-Fresco (2020), for a study methodology to become useful in the institutional context targeted by this study but also in the subtitling industry, it needs to be transferable: that is to say, *"straightforward, flexible and valid for training"*. The necessity of combining these needs with those of rigor certainly implied difficult decisions, as one of the main intentions was also that of simplifying the elements of the taxonomic and annotation model to make it more accessible for external evaluators, without compromising its rigor. The decision was that of favouring a simple taxonomy and annotation organization as complex annotation methods could prove too complicated or time-consuming for a potential subtitling company or institutional organization willing to replicate this strategy. The straightforward taxonomic scheme so defined significantly reduced the number of error classifications or the levels of severity, allowing replicating and transferring the system across different organizations or institutional situations. By contrast, other scholars, for example Eugeni (2008), preferred to promote a more complex taxonomic model which could determine and identify more detailed causes and types of errors in live subtitling. However, a wide array of error classifications might actually hinder the understanding of the evaluation process. As it was seen before under this discussion, the taxonomic scheme or features offered by literature are often source of ambiguity and complex categorizations. A simple ASR + NMT evaluation model can indeed offer the advantage of being relatively easy to understand and this aspect is of utmost importance in the case of large-scale projects, where it is necessary to train a high number of evaluators. But potential detractors may highlight that also simpler models can prove too complicated for a daily practical usage. In the case of the WER and NER models, for example, provided that these models are both based on the comparison between the original audio and the subtitles and that both need a transcription of the source speech (Gold Standard) to be carried out and analyzed, significant efforts in terms of time and costs are required in making an efficient evaluation (Romero-Fresco, 2020). Finally, it should be remarked that this study methodology was also "straightforward" in the sense that it offered results which can be easily readable by part of other users or evaluators. As a matter of fact, according to Romero-Fresco (2020), the results which a model produces *"should be measurable and recognizable"*.

5. Conclusions

At the end of this study, a series of final considerations will be offered in order to verify if the hypotheses made in the Introduction were confirmed and the Research Questions defined in §3.2 (Chapter 3) were responded. More specifically, it will be verified if the pipeline defined (AT + ASR + NMT system) in Chapter 3 can effectively help in breaking down the barriers of communications, while achieving a satisfactory accurate output as defined in this study.

First of all, with reference to the methodology adopted, it is possible to assert that the implementation of a statistical, quantitative approach provided an effective strategy in measuring the accuracy of the ASR + NMT system in generating automatic subtitles (without human intervention) in the specific scenario of this study, *i.e.*, conferences on climate change held at international organizations by Native and Non-Native speakers (in the mono-speaker mode). In particular, it is possible to claim that the statistical analysis and the implementation of the WER, NER and NTR models adopted here proved to be effective in measuring the accuracy of ASR-generated and NMT-translated subtitles. Yet, together with other scholars (Romero-Fresco and Pockhaker, 2017; Dawson, 2019), it is possible to maintain that the WER model was not adequate to appropriately measure the final output quality in terms of accuracy. In fact, as seen in Chapter 4, the model has the disadvantage that all errors (Substitutions, Deletions and Insertions) bear the same weight on the calculation of accuracy. Hence the necessity of adopting the NER statistical model in this study as the main "tool" of measurement emerged. Though it should not be considered as the only possible tool for an evaluation of accuracy, yet it is possible to maintain that the methodology implemented contributed to reach a reliable and possibly objective measurement of accuracy. De facto, the setting up of an Inter-Annotator Agreement test and the quite satisfying results obtained in terms of average agreement rates for the three taxonomic schemes adopted (as described in §4.3) permit to claim that this study deployed a reliable, effective and reproducible system of evaluation with average agreement rates well above 80%, described in more detail in §4.3. This however should not prevent from adopting other qualitative tools such as quality evaluation questionnaires or direct interviews to final users (as discussed in §4.9) in future studies to complement the evaluation of final output.

The surveying and ascertaining of alternative solutions to meet the increasing demand for subtitles evaluation should thus be continued and carried out not only in scientific literature, but also at an institutional level. On the other hand, it is necessary to add that, in a scenario like the one examined here, where AI (Artificial Intelligence) is of major importance, it should be underlined that the statistical approach can better help in assessing the quality of high-volume AI technology's output if compared to other methodologies (as also commented in Romero-Fresco 2011, 2015; Dawson, 2019). Finally, with respect to other studies on Institutional Translation, this empirical study probably contributed to achieve two goals: firstly, it possibly widened the observation perspectives on the multi-faceted, yet unexplored scenarios of translating and interpreting in the institutional contexts where AI technology is implemented; secondly, it eventually contributed to the collection of potential reusable and sharable data, thus encouraging comparison studies and follow-up analyses.

As far as the results of the analysis are concerned (see Chapter 4), it is possible to maintain that, across the entire ASR+NMT pipeline, the overall quality of the subtitles examined in this study was evaluated as sufficiently accurate for the Native speaker files only. In particular, quality was measured in terms of accuracy, which was examined in view of two different applications and usages: 1. Accuracy evaluation for intralingual subtitling for non-hearing people or people with a partial loss of hearing, and 2. Accuracy evaluation for interlingual subtitling into Italian (with the application of automatic Neural Machine Translation). For intralingual accuracy evaluation, in the case of VoxSigma-generated transcriptions, accuracy was well below the minimum accuracy rate (98%) set by the industry and defined in literature as seen in §4.5 when examining Non-Native speaker files; on the other hand, when considering the Native speaker files, the accuracy almost approached the minimum accuracy requisite with NER2 rate, i.e. when minor errors are excluded. In the case of GSR engine transcriptions for the sample examined, as described in §4.6, accuracy was again well below the minimum accuracy requisite (even if performing slightly better), except for the Native speaker files, where the software almost approached and overcame the threshold with NER1 and NER2 rates, respectively. For intralingual communication purposes, it should therefore be concluded that, with both groups of speakers (Native and Non-Native) under this study, the ASR technology, though responding to all the technological requisites seen in Chapter 2 (§2.2) and Chapter 3 (§3.4.1), actually failed to effectively meet the minimum accuracy rate. Yet, by taking into consideration the fact that the accuracy rate was mostly determined by Not Serious errors in the case of Native speakers, as seen in §4.4.3 of the analysis (Chapter 4), it is possible to conclude that with NER2 rate, both software solutions succeeded in meeting the industry's predefined threshold for accuracy. Overall, this general evaluation may also offer useful hints and evaluation considerations for the usage of ASR technology in different scenarios by part of respeakers in the production of live subtitling for non-hearing people. The NER rate was broken down into NER1 and NER2 rates in order to better represent the severity differentiation of errors, as well as to respond more efficaciously to the various applications of live subtitling (interlingual and intralingual subtitling for non-hearing people). In this respect, it may be tentatively suggested to use the NER2 rate for the evaluation of Native speaker files so as to eliminate the impact of minor errors (mainly Disfluency and Prosody related errors) in the calculation of accuracy.

However, for intralingual subtitling purposes in the present study's source language (English), it is plausible to maintain that the files having achieved WER and NER1 accuracy rates around 90% can be considered to be acceptable if human intervention is provided in the process of editing (respeaking process), including simultaneous editing of subtitle units, as claimed by Romero-Fresco (2016: 59). On the other hand, in the case of intralingual subtitling for people with a partial loss of hearing or with minor hearing difficulties, the situation would be different. In fact, these 90%-range accuracy subtitles could be considered to be understandable and usable for the final users, who are anyway capable of carrying out the lip reading technique at a conference setting in a live situation or who might have a partial hearing capacity (for example, old people). These subtitles would therefore represent an additional instrument for the breaking down of barriers in communications at an intralingual level. Yet the present study does not offer scientific grounds to confirm this final hypothesis. As a matter of fact, this would also depend on where the subtitles are made available either on a screen behind the speaker, or on a separate screen away from the speaker, or on the TV or computer screen where the event is broadcast. Additionally, for the purposes of intralingual subtitling (English) but addressed to totally non-hearing or deaf people, as well as for the purposes of interlingual subtitling into Italian (with the application of Neural Machine Translation), the subtitles generated from Non-Native speakers cannot be considered as sufficiently acceptable in terms of accuracy and it is therefore possible to conclude that the ASR+NMT technology examined here cannot provide for satisfactory results when a speaker is Non-Native. As a matter of fact, for interlingual subtitles in Italian, only the transcription files reaching an approximate accuracy rate of 98% with NER1 rate were treated under this study (as described in §4.6, Chapter 4).

To complete the conclusions on accuracy, it is possible to underline that, when comparing Google Speech Recognition (GSR)'s output with that of VoxSigma (VXS), the former proved to offer a slightly higher accuracy rate for the files examined (see §4.6, Chapter 4, for further details). More specifically, the aggregate percentage increase in accuracy amounted to about 1.3-1.5%. This improvement rate in terms of accuracy should be considered as particularly relevant for the selection of the appropriate software solution in the possible configuration of an ASR system for live subtitling at public conferences or for future studies. Yet it should be underlined that the AT feature commented in §4.8 would be available for the GSR solution only via Descript interface.

As far as interlingual subtitles in Italian are concerned, as already said above, it should be highlighted that only highly accurate transcriptions were submitted to the application of Automatic Machine Translation and the results obtained by measuring accuracy through the NTR model (§4.6, Chapter 4) were all above the minimum accuracy rate. In particular, the study calculated this rate only for a limited number of files. With the sample files examined (with Native speakers), the NTR rate was around or slightly above the 98% rate (as suggested in literature and required by the industry of subtitling for non-hearing people and for the purposes of multimedia accessibility), with a mean value of about 98.33%.

Under these conclusions, as hypothesized at the beginning of this study, another important consideration regards the innovative approach in considering the importance of terminology in the evaluation of accuracy. This study in fact showed that, with the application of Augmented Terminology resources, an innovative, effective strategy can be defined. As a matter of fact, by defining a new concept of "Augmented Terminology" and with the expansion of the ASR system built-in vocabulary, it was possible to establish a new AT+ASR+NMT pipeline based on Augmented Terminology, as described more in detail in §4.8. Additionally, this new concept finally brought to the proposal of defining an adapted version of the NER model based on a terminology categorization of errors. In the test conducted on a few files from the database, subtitles generated from the Native group of files were *de facto* significantly improved in terms of accuracy, and therefore they were enhanced for both intraligual and interlingual applications, contributing to breaking down the barriers of communication and automatic translation into the target language. As seen in §4.3.2, in general, the impact of Terminology-related errors was estimated around 16% (for Non-Native speaker files) and around 17% (for Native speaker files), with respect to all other error categories. In particular, by enhancing the terminological resources on an *a priori* basis, and by leveraging the validated termbase resources provided by the Food and Agriculture Organization (FAO) of the United Nations, it was possible to increase the accuracy of the ASR output.

To conclude, in addition to the implementation of AI in both ASR and NMT processes, one of the main results of this study was certainly the adoption of a combined approach for the analysis and evaluation of accuracy for subtitles. This approach was based, as seen above, on a statistical, quantitative model and also on a new concept of Augmented Terminology for the expansion of the built-in vocabulary of the ASR engine. This may probably contribute to the formulation of a new AT-adapted NER model based on terminology error categorization in future studies. In fact, as initially suggested in this study's Introduction (Chapter 1), the increasing demand for institutional translation at international organizations and the necessity of meeting the requirements of accessibility for physically-impaired people (hard of hearing people and non-hearing people or people with minor hearing difficulties) as provided by the EU Directive on Media Accessibility and other international legislation and standards, should be considered as stimuli for further investigations on ASR and subtitling for breaking down the barriers of communication.

Bibliography

Accipio Consulting (2006). “Tecnologie del linguaggio per l'Europa”, Report on the Technologies of language financed by the European Union. Available at: http://www.tcstar.org/pubblicazioni/ITC_ita.pdf

Adda-Decker, M. and Lamel, L. (2005). “Do speech recognizers prefer female speakers?”. Conference: *Interspeech 2005 - Eurospeech*, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4th-8th, 2005.

Al-Aynati, M. and Chorneyko, K. A. (2003). “Comparison of Voice-automated Transcription and Human Transcription in Generating Pathology Reports”, *Archives of Pathology and Laboratory Medicine*, 127(6): 721-725.

Alhawiti, K. M. (2015). “Advances in Artificial Intelligence Using Speech Recognition”. International Journal of Computer, Electrical, Automation, Control and Information Engineering. Vol:9. No:6.

Álvarez, A., Aliprandi, C., Gallucci, I., Piccinini, N., Raffaelli, M., del Pozo, A., Cassaca, R., Neto, J., Mendes, C. and Viveiros, M. (2015). “Automating Live and Batch Subtitling of Multimedia Contents for Several European Languages”, *Multimedia Tools and Applications* 75(18): 10823-10853.

Anderson, N. (2006). "Defense Department funds massive speech recognition and translation program". URL address: https://arstechnica.com/information-technology/2006/11/8186/.

Angelelli, C. V. (2004). Medical Interpreting and Cross-cultural Communication. Cambridge: Cambridge University Press.

Antonini, R., Cirillo, L., Rossato, L. and Torresi, I. (eds.) (2017). Non-professional Interpreting and Translation, Amsterdam/Philadelphia: John Benjamins.

Anusuya, M. A. and Katti, S. (2011). *“Front end analysis speech recognition: a review”*, In Int. J. Speech Technology, vol. 14, no. 2, 2011, pp. 99-145.

Apone, T., Brooks, M., and O’Connell, T. (2010). “Caption Accuracy Metrics Project. Caption Viewer Survey: Error Ranking of Real-time Captions in Live Television News Programs”. The WGBH National Center for Accessible Media. December 2010, Boston.

Armstrong, S. (1997). “Corpus-based methods for NLP and translation studies”. In Interpreting 2/1-2, pp. 141-162.

Artstein, R. and Poesio, M. (2008). “Inter-Coder Agreement for Computational Linguistics”. In Computational Linguistics. 34. 555-596.

Bachman, L., and Palmer, A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In *Proceedings of ICLR 2015*, 2015, San Diego, California, USA. Available online: https://bit.ly/2nZ3IAc.

Bassnett-McGuire, S. (1991). *Translation Studies*. London & New York: Routledge.

Beaufays, F. (2015). *The neural networks behind Google Voice transcription*. In Google AI Blog. https://research.googleblog.com/2015/08/the-neural-networks-behind-google-voice.html

Bendazzoli, C. (2010). “Corpora e interpretazione simultanea”. Bologna: Asterisco, p. 264. DOI 10.6092/unibo/amsacta/2897. In: Alma-DL. Saggi

Bendazzoli, C. and Sandrelli, A. (2005). “An Approach to Corpus-Based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus)”. MuTra 2005 – Challenges of Multidimensional Translation: Conference Proceedings.

Bentivogli, L.,Bisazza, A., Cettolo, M. and Federico, M. (2016). “Neural *versus* Phrase-Based Machine Translation Quality: a Case Study”. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, Austin, Texas. 257–267. Available online: https://bit.ly/2nZnP1b.

Bernardini, S., Ferraresi, A., Milicevic, M. (2016). “From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective”. In *Target* (28): 61-86

Bernardini, S., Ferraresi, A., Russo, M., Collard C. and Defrancq, B. (2018). “Building Interpreting and Intermodal Corpora: A How-to for a Formidable Task”. In book: *Making Way in Corpus-based Interpreting Studies*. Series: New Frontiers in Translation Studies. Russo, M., C. Bendazzoli & B. Defrancq (Eds.). XVI, 215 pages. Springer Singapore

Bersani Berselli, G. (ed.) (2011). *Usare la traduzione automatica*. Bologna: CLUEB.

Besnier, J.-M. (2012). *L’homme simplifié: Le syndrome de la touche étoile*. Paris: Fayard.

Bettinson, M. (2013) *The Effect of Respeaking on Transcription Accuracy*, Unpublished Honours Thesis, Melbourne: University of Melbourne.

Bowker, L. (2000). A Corpus-Based Approach to Evaluating Student Translations. *The Translator*, *6*(2), 183–209.

Branchadell, A. and West, L.M. (eds.) (2005), Less translated languages, Amsterdam/Philadelphia: John Benjamins.

Braun, S, Davitti, E., and Dicerto, S. (2016). AVIDICUS 3 Project - Research Report. University of Surrey.

Britannica, The Editors of Encyclopedia. "Database". *Encyclopedia Britannica*, 18 May. 2020, https://www.britannica.com/technology/database. Accessed: 13 May 2021.

Brunette, L. (2000). Towards a Terminology for Translation Quality Assessment: A Comparison of TQA Practices. *The Translator*, *6*(2), 169-182.

Caimi, A. (2006). “Audiovisual Translation and Language Learning: The Promotion of Intralingual Subtitles”. In *The Journal of Specialized Translation*. 6.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Way A. and Georgakopoulou, P. (2018a). “Evaluating MT for massive open online courses: A multifaceted comparison between PBSMT and NMT systems”. *Machine Translation*, 32: 225-278.

Castilho, S., Doherty, S., Gaspari, F. and Moorkens, J. (2018b). “Approaches to Human and Machine Translation Quality Assessment”. In J. Moorkens, S. Castilho, F. Gaspari & S. Doherty (2018). *Translation quality assessment: from principles to practice.* 9-38.

Castilho, S., Gaspari, F., Moorkens, J., Popovic, M. and Toral, A. (2019). Editors’ foreword to the special issue on human factors in neural machine translation. Machine Translation. Project: MT journal Special Issue on Human Factors in Neural Machine Translation. 33. DOI: 10.1007/s10590-019-09231-y

Chang, C.-C, Wu, M, M-C and Kuo, T.-C. G. (2018). *“Conference interpreting and knowledge acquisition: How professional interpreters tackle unfamiliar topics”*. In *Interpreting* 20(2). 204-231.

Chiari, I. (2007). Introduzione alla linguistica computazionale. GLF Editori Laterza

Chiari, I. (2011). “Traduzione automatica”. *X la Tangente*, aprile, 31-33.

Cho, K., Van Merriënboer, B., Bahdanau, D. and Bengio, Y. (2014). “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches”. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014, Doha, Qatar. 103-11. Available online: https://bit.ly/2o54xY5.

Clifford, A. (2001). Discourse Theory and Performance-Based Assessment: Two Tools for Professional Interpreting. *Meta: Journal Des Traducteurs*, *46*(2), 365-378.

Cockhaert, H.J, and Steurs, F. (2015). *Handbook of Terminology*. John Benjamins Publishing Company Amsterdam/Philadelphia.

Cogo, A. and Jenkins, J. (2010). “English as a Lingua Franca in Europe. A mismatch between policy and practice”. In *European Journal of Language Policy*. October 2010: 271-293

Corpas Pastor, G. and Fern, L. M. (2016). *A survey of interpreters’ needs and practices related to language technology*. Tech. rep. Málaga: University of Málaga.

Crystal, D. and Robins, R.H. (2020). "Language". *Encyclopedia Britannica*, 29 Oct. 2020, https://www.britannica.com/topic/language. Accessed 24 January 2021.

Davitti, E. and Sandrelli, A. (2020). “Embracing the Complexity: a pilot study on Interlingual Respeaking”. In Journal of Audiovisual Translation; European Association for Studies in Screen Translation.

Dawson, H. (2019). “Feasibility, quality and assessment of interlingual live subtitling: A pilot study”. Journal of Audiovisual Translation, 2(2), 36-56.

De Wachter, M., Matton, M., Demuynck, K., Wambacq, P, Cools, R. and Van Compernolle, D. (2018). "Template-Based Continuous Speech Recognition," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1377-1390, May 2007,

Deng, L. and Yu, D. (2014). "Deep Learning: Methods and Applications", Foundations and Trends in Signal Processing: Vol. 7: No. 3-4, pp 197-387.

Deng, L. and Li, X. (2013). *"Machine Learning Paradigms for Speech Recognition: An Overview"*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 21 no. 5, 2013, pp.1060-1089.

Desmet, B., Vandierendonck, M. and Defrancq, B. (2018). "Simultaneous interpretation of numbers and the impact of technological support". In Claudio Fantinuoli (ed.), *Interpreting and technology*, 13–27. Berlin: Language Science Press.

D'Hayer, D. (2012). Public Service Interpreting and Translation: Moving Towards a (Virtual) Community of Practice. Meta: Journal des traducteurs. 57. 235. 10.7202/1012751ar.

Doherty, S. (2017). "Issues in human and automatic translation quality assessment". In Dorothy Kenny (Ed.), *Human Issues in Translation Technology* (pp. 131–148). London: Routledge.

Doherty, S., Gaspari, F., Groves, D., and Van Genabith, J. (2013). Mapping the Industry I: Findings on Translation Technologies and Quality Assessment.

Dumouchel, P., Boulianne, G. and Brousseau, J. (2011). "Measures for Quality of Closed Captioning". In A. Şerban, A. Matamala and J.-M. Lavaur (eds) *Audiovisual Translation in Close-up: Practical and Theoretical Approaches*, Bern: Peter Lang, 161-172.

Dureja, M and Gautam, S. (2015). "Speech-to-Speech Translation: A Review". In International Journal of Computer Applications (0975 – 8887). Volume 129 – No.13, November 2015.

Eddy, S. (2004). *"What is a hidden Markov model?"*. Nat Biotechnol 22, 1315-1316

Errattahi, R., El Hannani, A. and Ouahmane, H. (2018). "Automatic speech recognition errors detection and correction: A review". Procedia Computer Science 128, 32-37.

Errattahi, R., El Hannani, A., Ouahmane, H. and Hain, T. (2016). "Automatic speech recognition errors detection using supervised learning techniques". 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)

Eugeni, C. (2008). "La sottotitolazione in diretta TV. Analisi strategica del rispeakeraggio verbatim di BBC News". PhD Thesis, University of Naples.

Eugeni, C. (2009). "Respeaking the BBC news: a strategic analysis of respeaking on the BBC". In The Sign Language Translator and Interpreter (SLTI), 3 (1), 29-68

European Union (2020). *Interinstitutional Style Guide*. Publication Office of the European Union. Available on: https://publications.europa.eu/code/en/en-000100.htm (Last visited on 17/12/2020)

Fairclough, N. (2013). Critical discourse analysis: The critical study of language (2^{nd} ed.). London: Longman.

Falletto, A. (2007). "*Che cosa è, come funziona: Algoritmi e tecnologie per il riconoscimento vocale". Stato dell'arte e sviluppi futuri*. http://www.crit.rai.it/eletel/2007-2/72-6.pdf

Fantinuoli, C. and Prandi, B. (2018). *"Teaching information and communication technologies. A proposal for the interpreting classroom"*. In Trans-Kom 11 [2] (2018): 162-182.

Fantinuoli, C. (2016). *"InterpretBank. Redefining computer-assisted interpreting tools"*. In Proceedings of the Translating and the Computer 38 Conference, 42-52, London. Editions Tradulex.

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

Fantinuoli, C. (2017b). *"Speech Recognition in the Interpreter Workstation"*. Conference: Translating and the computer 3, London, Project: Information Technology and Interpreting, November 2017.

Fantinuoli, C. (2018a). "*Interpreting and technology: The upcoming technological turn"*. In Fantinuoli, C. (eds) *Interpreting and Technology* (Translation and Multilingual Natural Language Processing 11). Berlin: Language Science Press, 2018.

Fantinuoli, C. (2018b). *"Computer-assisted interpreting: Challenges and future perspectives"*. In Gloria Corpas Pastor & Isabel Durán-Muñoz (eds.), *Trends in E-tools and resources for translators and interpreters*, 153-174. Leiden: Brill.

Fosler-Lussier, E. and Morgan, N. (1999). "Effects of speaking rate and word frequency on pronunciations in conversational speech". In Speech Communication. Volume 29, Issues 2–4, 1999. 137-158.

Freitag, M., Peitz, S., Wuebker, J., Ney, H., Durrani, N., Huck, M., Koehn, P., Ha, T., Niehues, J., Mediani, M., Herrmann, T., Waibel, A., Bertoldi, N., Cettolo, M. and Federico, M. (2013): "EU-BRIDGE MT: text translation of talks in the EU-BRIDGE project". [IWSLT 2013] Proceedings of the 10th International Workshop on Spoken Language Translation, Heidelberg, Germany, Dec. 5-6, 2013; 8 pp.

Fu, K.S. (eds) (1976). "*Introduction"*. In *Digital Pattern Recognition*. Communication and Cybernetics, vol 10. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-96303-2_1

Fügen, C., Kolss, M., Bernreuther, D., Paulik, M., Stuker, S., Vogel S. and Waibel, A. (2006). "Open Domain Speech Recognition & Translation: Lectures and Speeches". Conference: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on Volume: 1. DOI: 10.1109/ICASSP.2006.1660084

Fuoli, M. and Hommerberg, C. (2015). "Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions". In Corpora. 10. 315-349. 10.3366/cor.2015.0080.

Gada J., Rao, P. and Samudravijaya, K. (2013). “Confidence Measures for Detecting Speech Recognition Errors”. In Proceedings of National Conference on Communications, 15-17 February, 2013, New Delhi.

Gagliardi, G. (2018). “Inter-Annotator Agreement in linguistica: una rassegna critica”. Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin.

Garofalo, J. S., Lamel, L.F, Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L. (1993). “DARPA TIMIT acoustic phonetic continuous speech corpus”. CD-ROM, Tech. rep., U.S. Dept. of Commerce, NIST, Gaithersburg, MD.

Garofolo, J. S., Laprun, C.D., Michel, M., Stanford V.M. and Tabassi, E. (2004). “The NIST Meeting Room Pilot Corpus”. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.331.5983&rep=rep1&type=pdf

Gaspari, F. (2011). “Introduzione alla traduzione automatica”. In G. Bersani Berselli (2011). 13-31.

Gaspari, F. and Hutchins, J. (2007). “Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects”. *Proceedings of Machine Translation Summit XI*, Copenhagen Business School, Copenhagen (Denmark), 10-14 September 2007, 199-206.

Gazzola, M. (2016). “Research for cult committee. European strategy for multilingualism: benefits and costs”. Brussels: European Union. http://www.europarl.europa.eu/RegData/etudes/STUD/2016/573460/IPOL_STU(2016)573460_EN.pdf

Ghai, W. and Singh, N. (2012). *“Literature Review on Automatic Speech Recognition”*. International Journal of Computer Applications. 41. 42-50. 10.5120/5565-7646.

Gile, D. (2009). *Basic Concepts and Models for Interpreter and Translator Training*: Revised edition. John Benjamins Publishing Company, Amsterdam, 2nd edition.

Goldwater, S., Jurafsky, D. and Manning, C.D. (2010). “Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates”. Speech Communication 52 (2010), 181-200.

Goman, A. (2017). “Addressing Hearing Loss With an Aging Population”, The Hearing Journal: June 2017 - Volume 70 - Issue 6: 6.

González Núñez, G. (2014), Translating for linguistic minorities: translation policy in the United Kingdom, Universitat Rovira I Virgili, PhD Thesis, https://bit.ly/2CfqnuX.

Gordon, G. N. (2020). "Communication". *Encyclopedia Britannica*, 16 Dec. 2020, https://www.britannica.com/topic/communication. Accessed 24 January 2021.

Gouadec, D. (2007). *Translation as a Profession*. Amsterdam/Philadelphia: John Benjamins.

Greco, G. M. (2016). “On Accessibility as a Human Right, with an Application to Media Accessibility”. In A. Matamala and P. Orero (eds.), Researching Audio Description. New Approaches, Palgrave 2016, pp. 11-33.10.1057/978-1-137-56917-2_2.

Greco, G. M. (2018). The Case for Accessibility Studies. *Journal of Audiovisual Translation*, *1*(1).

Hemdal, J.F. and Hughes, G.W. (1967). *“A feature based computer recognition program for the modeling of vowel perception”*. In Models for the Perception of Speech and Visual Form*,* Wathen-Dunn, W. Ed. MIT Press, Cambridge, MA.

Hinton, G, Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B. (2012). *"Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups"*, IEEE Signal Process. Magazine, vol. 29, no. 6, (2012), pp. 82-97.

Hirschberg, J., Litman, D. and Swerts, M. (2004). “Prosodic and Other Cues to Speech Recognition Failures”. In Speech Communication, Volume 43, Issues 1-2 , June 2004, Pages 155-175.

Hochreiter, S. and Schmidhuber, J. (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735.

House, J. (2009). Quality. In *Routledge Encyclopaedia of Translation Studies:* 222-225. London & New York: Routledge.

Huang X. and Deng, L. (2010). “*An Overview of Modern Speech Recognition*”. In *Handbook of Natural Language Processing,* Second Edition, Chapter 15, Chapman & Hall/CRC, 2010, pp. 339-366.

Huang, X, Acero, A. and Hon, H. W. (2001). “Spoken Language Processing: a guide to theory, algorithm, and system development”, Prentice Hall, 2001.

Hutchins, J. (2005). “Example-based Machine Translation: a Review and Commentary”. *Machine Translation*, 19: 197-211. Available online: https://bit.ly/2pxNsGt.

Hutchins, J. and Somers, H. L. (1992). *An introduction to machine translation.* Londra: Academic Press.

Hutchins, W. J. (1995). “Machine Translation: a brief history”. In *Concise history of the language sciences: from the Sumerians to the cognitivists*. Edited by E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press, 1995. Pages 431-445

Hutchins, W. J. (2010). “Machine translation: a concise history”. Journal of Translation Studies, 13, 1-2: 29-70. Also available online: http://www.hutchinsweb.me.uk/CUHK-2006.pdf

Hutchins, W. J. (eds.) (2000). “Early Years in Machine Translation: Memoirs and Biographies of Pioneers”. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Hutchins, W. J. and Somers, H. L. (1992). “An introduction to machine translation”. Academic Press, London.

ISPI, Istituto per gli Studi di Politica Internazionale (2012). Report "The long walk to gender parity in international organizations". Publications for the Italian Parliament and Ministry of Foreign Affairs. Published on 12th July, 2012. Available at: https://www.ispionline.it/it/pubblicazione/long-walk-gender-parity-international-organizations

Itakura, F. (1975). "Minimum prediction residual principle applied to speech recognition". IEEE Trans. on Acoustics, Speech, and Signal Processing (ASSP), ASSP-23(1): 67-72, February 1975.

Jakobson, R. (1959). "On linguistic aspects of translation", in Brower, R.A. (eds.) *On Translation*. Cambridge, MA: Harvard University Press, 232-239.

Jenkins, J. (2000). *The Phonology of English as an International Language* (Oxford: Oxford University Press).

Jenkins, H. (2009). "Confronting the Challenges of Participatory Culture". Media Education for the 21st Century. Cambridge. 145 pages. URI: http://library.oapen.org/handle/20.500.12657/26083

Jopek Bosiacka, A. (2013). "Comparative law and equivalence assessment of systembound terms in EU legal translation". In: *Linguistica Antverpiensia*, New Series – Themes in Translation Studies, No 12 (2013). Link: https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/237

Jurafsky, D. and Martin, J. H. (2009). *"Speech and Language Processing"*. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, 2009.

Kachru, B. B. (1985). "Standards, codification and sociolinguistic realism: the English language in the outer circle". In *English in the World: Teaching and Learning the Language and Literatures*, Cambridge: Cambridge University Press, 1985, pp. 11-30.

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

Kalina, S. (2005). *"Quality Assurance for Interpreting Processes"*. Meta: Journal des traducteurs, 50(2): 768.

Kang, J.H. (2012). "Institutional Translation". Routledge Encyclopedia of Translation Studies 2nd ed. by Mona Baker; Gabriela Saldanha. Machine Translation. 26: 271-275.

Kang, J.H. (2014). "Institutions translated: discourse, identity and power in institutional mediation". In Perspectives, 22:4: 469-478.

Karpagavalli, S. and Chandra, E. (2016). "A Review on Automatic Speech Recognition Architecture and Approaches". International Journal of Signal Processing, Image Processing and Pattern Recognition. 9: 393-404.

Kay, M., J. M. Gawron, & P. Norvig (1994). "Verbmobil: A Translation System for Face-to-Face Dialog". Stanford, CSLI Lecture Notes No. 33 CA: CSLI Publications.

Kenny, D. (2018). “Sustaining Disruption? The Transition from Statistical to Neural Machine Translation”. *Revista Tradumàtica*. 16: 59-70

Klubička, F., Toral, A. and Sánchez-Cartagena, V. M. (2017). “Fine-Grained Human Evaluation of Neural Versurs Phrase-Based Machine Translation”. *The Prague Bulletin of Mathematical Linguistics*, 108: 121-132. Available online: https://bit.ly/2oa75US.

Koehn, P. (2009). *Statistical machine translation.* Cambridge: Cambridge University Press.

Koponen, M. (2012). Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations. In *Proceedings of the 7th Workshop on Statistical Machine Translation* (pp. 181–190). Montreal, Canada.

Koskinen, K. (2000). “Institutional Illusions”. In *The Translator*, 6:1: 49-65, DOI: 10.1080/13556509.2000.10799055

Koskinen, K. (2008). “Translating Institutions. An Ethnographic Study of EU Translation”. Manchester/Kinderhook: St. Jerome Publishing.

Koskinen, K. (2014). “Institutional translation: the art of government by translation”. *Perspectives*, 22:4: 479-492, DOI: 10.1080/0907676X.2014.948887

Kress, G. (1995). “The social production of language: History and structures of domination”. In P. Fries & M. Gregory (Eds.), Discourse in society: Systemic functional perspectives: 169-191. Norwood, NJ: Ablex.

Kumar, R., Hewavitharana, S., Zinovieva, N., Roy, M. E. and Pattison-Gordon, E. (2015). “Error-tolerant speech-to-speech translation”. MT Summit XV, October 30th–November 3rd, 2015, Miami, Florida, USA. Proceedings of MT Summit XV: vol.1: MT Researchers’ Track; 229-239.

Lazzari, G. (2006). "TC-STAR: a speech to speech translation project", In International Workshop on Spoken Language Translation (IWSLT) 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006.

Lederer, M. (1978). “Simultaneous Interpretation - Units of Meaning and other Features”. In *Language Interpretation and Communication*: 323-332.

Lewis, W. (2015). “Skype Translator: breaking down language and hearing barriers. A behind the scenes look at near real-time speech translation”. Proceedings of the 37th Conference Translating and the Computer, London, November 26th-27th, 2015; 58-65.

Li, J, Deng, L, Gong, Y and Umbach, R. H. (2014). “An Overview of Noise-Robust Automatic Speech Recognition”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, 2014: 745-777.

Liu, Q. and Zhang, X. (2014). "Machine Translation". In The Routledge Encyclopaedia of Translation Technology ed. Chan Sin-wai (Abingdon: Routledge, 03 Nov. 2014), Routledge Handbooks Online.

Liu, S., Hu, S., Liu, X. and Meng, H. (2019). “On the Use of Pitch Features for Disordered Speech Recognition”. Proceedings Interspeech 2019, September 15th-19th, 2019, Graz, Austria. Available online: https://www.isca-speech.org/archive/Interspeech_2019/pdfs/2609.pdf

Luce, P. A. and Pisoni, D. B. (1998). “Recognizing spoken words: the neighbourhood activation model”. In *Ear and hearing*, 19(1): 1-36.

Maffi, A. (2016). “Studio e sviluppo di un framework per il riconoscimento vocale nell’ambito di sistemi hands-free”. Degree Thesis. University of Bologna.

Martín Ruano, M. R. (2014). “From Suspicion to Collaboration: Defining New Epistemologies of Reflexive Practice for Legal Translation and Interpreting”. In The Journal of Specialised Translation. 22.

Maslias, R. (2017). “In Termino Qualitas. In Human and Machine Translation”. First World Congress on Translation Studies. Workshop on Computer Assisted Translators vs. Human Translation. Paris, Nanterre University, 11th-12th April 2017.

Matusov, E., Leusch, G., Banchs, R., Bertoldi, N., Dechelotte, D., Federico, M., Kolss, M., Lee, Y.S., Marino, J., Paulik, M., Roukos, S., Schwenk, H. and Ney. H. (2008). “System Combination for Machine Translation of Spoken and Written Language”. IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 7: 1222-1237.

Matusov, E., Ueffing, N. and Ney, H. (2006). “Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment,” in Conference of the European Chapter of the Association for Computational Linguistics (EACL): 33-40.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, M., Masson, A., Post, W., Reidsma D. and Wellner, P. (2005). “The AMI meeting corpus”. Fifth International Conference on Methods and Techniques in Behavioral Research, 30th August-2nd September 2005, Wageningen, The Netherlands.

McEnery, T. and Wilson, A. (1996). Corpus Linguistics. Edinburgh University Press. 209 pages.

McKean, K (1980). “When Cole talks, computers listen”. Sarasota Journal, April 8th, 1980.

Meylaerts, R. (2010), “Multilingualism and translation”, in Yves Gambier and Luc van Doorslaer, Handbook of Translation Studies. Amsterdam/Philadelphia: John Benjamins, vol. I: 227-230.

Meylaerts, R. (2011), “Translational Justice in a Multilingual World: An Overview of Translational Regimes”, in Meta, vol. LVI, n. 4: 743-757. doi:10.7202/1011250ar.

Mirzaei, M. S., Meshgi, K. and Kawahara, T. (2018). “Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening”. In Computer Speech & Language. Volume 49, 2018. Pages 17-36.

Modiano, M. (2017). “English in a post-Brexit European Union”. In World Englishes. 36(1).

Mossop, B. (1990). *"Translating Institutions and 'Idiomatic' Translation".* Revised version of a paper originally published in Meta. Translators' Journal http://www.yorku.ca/brmossop/TranslatingInstitutionsRevised.htm (consulted 16/12/2020).

Mostefa, D., Hamon, O. and Choukri, K. (2006). "Evaluation of Automatic Speech Recognition and Speech Language Translation within TC-STAR: Results from the first evaluation campaign".

Müller, M. (2007). *"Dynamic Time Warping".* In: Information Retrieval for Music and Motion. Springer, Berlin, Heidelberg.

Müller, M., Nguyen, T. S., Niehues, J., Cho, E., Krüger, B., Ha, T.L., Kilgour, K., Sperber, M., Mediani, M., Stüker, S. and Waibel, A. (2016). "*Lecture Translator: Speech translation framework for simultaneous lecture translation".* 82–86. Association for Computational Linguistics.

Nakamura, S. (2009). "Overcoming the Language Barrier with Speech Translation Technology". *Science & Technology Trends -Quarterly Review*, n. 31, April 2009, 36-48.

Naldi, M. (2014). *Traduzione automatica e traduzione* assistita. Bologna: Esculapio.

Neves, J. (2018). *Subtitling for Deaf and Hard-of-hearing Audiences: Moving Forward*. (Luis Pérez-González, Ed.), *The Routledge Handbook of Audiovisual Translation*. London: Routledge.

O'Shaughnessy, D. (2008). Invited paper: "Automatic speech recognition: History, methods and challenges", Pattern Recognition, Volume 41, Issue 10, 2008: 2965-2979.

Ofcom (2015). *Measuring the quality of live subtitling*. London. Retrieved from https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/live-subtitling

Olive, J., Christianson, C. and McCary, J. (2011). Handbook of Natural Language Processing and Machine Translation. "DARPA Global Autonomous Language Exploitation". XXVI, 936. Springer-Verlag, New York.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation". In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, Philadelphia, Pennsylvania, USA. 311-318. Available online: https://bit.ly/2pIOoYZ.

Pedersen, J. (2017). "The FAR model: Assessing quality in interlingual subtitling". In Journal of Specialized Translation, 28: 210-229.

Peng, L. and Ann, J. (2000). "Stress and Duration in Three Varieties of English". In *World Englishes* 20.1: 1-27.

Phillipson, R. (2003). *English-Only Europe? Challenging Language Policy* (London: Routledge).

Pierce, J.R. (1969). "Whither speech recognition?". In *The Journal of the Acoustical Society of America,* 46: 1049.

Pitzl, M.-L., Breiteneder, A. and Klimpfinger, T. (2008). “A World of Words: Processes of Lexical Innovation in VOICE”, Vienna English Working Papers 17.2: 21-46.

Prandi, B. (2018). *“An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation”*. In Claudio Fantinuoli (ed.), *Interpreting and technology*, 29-59. Berlin: Language Science Press.

Rabiner, L. and Juang, B.H. (1993). *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall, Apr. 1993, vol. 14.

Rajadurai, J. (2005). “Revisiting the Concentric Circles: Conceptual and Sociolinguistic Considerations”. Asian EFL Journal, 7(4)

Ray, R., DePalma, D., and Pielmeier, H. (2013). *The Price-Quality Link*. US: Common Sense Advisory, Ltd.

Remael, A., Orero, P., and Carroll, M. (2012). *Audiovisual Translation and Media Accessibility at the Crossroads*. Amsterdam: Rodopi.

Riediger, H. and Galati, G. (2012).“Tecnologie per la traduzione”. Also available online: http://www.fondazionemilano.eu/moodle/course/view.php?id=278

Rindler-Schjerve, R., and Vetter, E. (2007). “Linguistic Diversity in Habsburg Austria as a Model for Modern European Language Policy”, in J. ten Thije and L. Seevaert (eds.). Receptive Multilingualism (Amsterdam: John Benjamins): 49-69.

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

Romero-Fresco, P. (2016). “Accessing communication: The quality of live subtitles in the UK”. In *Language & Communication*. 49: 56-69.

Romero-Fresco, P. (2015). *The Reception of Subtitles for the Deaf and Hard of Hearing in Europe*. Bern/Berlin/Bruxelles/Frankfurt am Main/New York/Oxford/Wien: Peter Lang.

Romero-Fresco, P. (2018). “In support of a wide notion of media accessibility: Access to content and access to creation”. *Journal of Audiovisual Translation*. 187-204.

Romero-Fresco, P. (2020). “Negotiating quality assessment in media accessibility: the case of live subtitling”. Universal Access in the Information Society. https://doi.org/10.1007/s10209-020-00735-6

Romero-Fresco, P. and Martínez, J. (2015). “Accuracy Rate in Live Subtitling: The NER Model”. In J. Díaz-Cintas and R. Baños (Ed.), *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape:* 28-50. London: Palgrave MacMillan.

Romero-Fresco, P. and Pöchhacker, F. (2017). “Quality assessment in interlingual live subtitling: The NTR model”. In *Linguistica Antverpiensia*, New Series: Themes in Translation Studies, 16, 149-167.

Ruiz, N. and Federico, M. (2014). “Assessing the Impact of Speech Recognition Errors on Machine Translation Quality”.

Ruiz, N., Di Gangi, M.A., Bertoldi, N. and Federico, M. (2017). “Assessing the Tolerance of Neural Machine Translation Systems Against Speech Recognition Errors”. Interspeech 2017.

Russo, M., Bendazzoli, C., Sandrelli, A. and Spinolo, N. (2012). “The European Parliament Interpreting Corpus (EPIC): Implementation and developments”. In *Breaking ground in corpus-based interpreting studies*. (eds.). F. Straniero Sergio, and C. Falbo, 35-90. Bern: Peter Lang.

Salimbajevs, A. and Strigins, J. (2015). “Error Analysis and Improving Speech Recognition for Latvian Language”. RANLP.

Samoulian, A. (1994). “Knowledge Based Approach to Speech Recognition”. Fifth Australian International Conference on Speech, Science and Technology.

Sandrelli, A and De Manuel Jerez, J. (2007). “The Impact of Information and Communication Technology on Interpreter Training State-of-the-art and Future Prospects”. In *The Interpreter and translator trainer* 1(2): 269-303.

Saon, G. and Chien, J. (2012). "Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances". In IEEE Signal Processing Magazine, vol. 29, no. 6: 18-33, Nov. 2012.

Schäffner, C., Tcaciuc, L.S. and Tesseur, W. (2014). “Translation practices in political institutions: a comparison of national, supranational, and nongovernmental organisations”. In *Perspectives*, vol. XXII, n. 4: 493-510.

Shinozaki, T. and Furui, S. (2001). “Error analysis using decision trees in spontaneous presentation speech recognition”. Conference: Automatic Speech Recognition and Understanding, 2001. ASRU '01: 198-201.

Sinclair, J. M. (1996). “EAGLES preliminary recommendations on text typology”. Available on: www.ilc.cnr.it/EAGLES/texttyp/texttyp.html. (Last visited on 16/12/2020).

Snell-Hornby, M. (1992). The Professional Translator of Tomorrow: Language Specialist of All-round Expert? In *Teaching Translation and Interpreting: Training, Talent and Experience:* 9-22. Amsterdam: John Benjamins.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). “A Study of Translation Edit Rate with Targeted Human Annotation”. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, 2006, Cambridge, Massachusetts, USA. 223-231. Available online: https://bit.ly/2pHjKPy.

Song, X., Cohn, T., and Specia, L. (2013). “BLEU Deconstructed: Designing a Better MT Evaluation Metric”. URL: https://www.semanticscholar.org/paper/BLEU-Deconstructed%3A-Designing-a-Better-MT-Metric-Song-Cohn/89eca547b1a2f6208ae529d45a65a51cf49adbff

Sperber, M., Neubig, G. and Fügen, C. (2013). “Efficient Speech Transcription through Respeaking”. In *InterSpeech*: 1087-1091.

Spinolo, N., Bertozzi, M. and Russo, M. (2018). "Shaping the Interpreters of the Future and of Today: Preliminary results of the SHIFT Project". In *The Interpreters' Newsletter 2018* (23): 45-61.

Spooren, W. and Degand, L. (2010). "Coding coherence relations: Reliability and validity". In *Corpus Linguistics and Linguistic Theory*. 6. 10.1515/cllt.2010.009.

Standardization, I. O. for. (2018). Information technology – user interface component accessibility – Part 23: Guidance on the visual presentation of audio information (including captions and subtitles) *(ISO/IEC DIS 20071-23: 2018)*. Retrieved from https://www.iso.org/standard/70722.html

Starnoni, E. (2019). "Traduttori umani e traduzione automatica neurale". *Il Chiasmo*. Available online: https://bit.ly/2m6pald.

Stein, D. (2018). Machine translation: Past, present and future. Language technologies for a multilingual Europe, 4(5).

Stuckless, R. (1994). "Developments in real-time speech-to-text communication for people with impaired hearing," in *Communication access for people with hearing loss*: 197-226.

Sturari, N. (2012). *Riconoscimento vocale e smartphone: sviluppo di un'applicazione capace di dialogo su piattaforma Android*. Dissertation Thesis. Ancona: Università Politecnica delle Marche. http://airtlab.dii.univpm.it/it/system/files/thesis/sturari-nicola-thesis.pdf

Sutskever, I., Vinyals, O. and Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks". In *Proceeding of NIPS 2014*, 2014, Montréal, Canada. Available online: https://bit.ly/2KWaPDw.

Thompson, P. (2005). "Spoken language corpora". In Wynne, M. (eds.). Online: http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter5.htm/ (Last visited on 16/12/2020).

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. Studies in corpus linguistics. Vol. 6. John Benjamins Publishing (eds.). 223 pages.

Toral, A., and Sánchez-Cartagena, V. M. (2017). "A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions". In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, Valencia, Spain. 1: 1063-1073. Available online: https://bit.ly/2n95hea.

Tripathy, H. K., Tripathy, B. K., and Das, P. K. (2008). *"A Knowledge based Approach Using Fuzzy Inference Rules for Vowel Recognition"*, Journal of Convergence Information Technology Vol. 3 No 1, March 2008.

Valentini, C. (2002). *Uso del computer in cabina di interpretazione.* http://aiic.net/page/attachment/960

Valor Miró, J. D., Turró, C., Civera, J. and Juan, A. (2015). "Evaluación de la revisión de transcripciones y traducciones automáticas de vídeos Polimedia", *Proceedings of I*

Congreso Nacional de Innovación Educativa y Docencia en Red (IN-RED 2015) Valencia (Spain): 461-465.

Van Brussel, L., Tezcan, A. and Macken, L. (2018). “A Fine-grained Error Analysis of NMT, PBMT and RBMT Output for English-to-Dutch”. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018, Miyazaki, Japan: 3799-3804. Available online: https://bit.ly/2niwj32.

Van Gerven, M. and Bohte, S. (eds.) (2018). Artificial Neural Networks as Models of Neural Information Processing. Research Topic. Frontiers. Available online: https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing

Vilar, D., Matusov, E., Hasan, S., Zens, R. and Ney, H. (2005). “Statistical machine translation of European parliamentary speeches”. In Proceedings of MT Summit.

Vilar, D., Xu, J., D'Haro, L. and Ney, H. (2006). “Error analysis of statistical machine translation output”. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06), Genoa, European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf

Vitevitch, M.S. and Luce, P.A. (1999). “Probabilistic phonotactics and neighborhood activation in spoken word recognition”. In *Journal of Memory and Language*. 1999; 40: 374-408.

Wahlster, W. (1993). “Verbmobil: Translation of Face-to-Face Dialogs”. In *Proceedings of the Fourth Machine Translation Summit*, Kobe, Japan: 128-135.

Wahlster, W. (2000). "Mobile Speech-to-Speech Translation of spontaneous dialogs: an overview of the final Verbmobil system". Verbmobil: Foundations of Speech-to-Speech Translation. Springer Berlin Heidelberg, 2000: 3-21.

Wahlster, W. (2001). “Robust Translation of Spontaneous Speech: A Multi-Engine Approach”. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, Seattle, Washington, August 2001, San Francisco: Morgan Kaufmann, Vol. 2: 1484-1493

Wahlster, W., (eds.) (2000). *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer, Berlin.

Way, A. (2018). “Quality Expectations of Machine Translation”. In J. Moorkens, S. Castilho, F. Gaspari and S. Doherty (2018). *Translation quality assessment: from principles to practice*: 159-178.

Weaver, W. (1949). Recent Contributions to the Mathematical Theory of Communication. Available at: http://www.panarchy.org/weaver/communication.html

White, R. (1997). “Going Round in Circles: English as an International Language, and Cross Cultural Capability”. *Proceedings of the Languages for Cross-cultural capability Conference*, Leeds Metropolitan University, 12th-14th, December.

Wu, Y, Schuster, M., Chen, Z., Le Q.V., Norouzi M. et al. (2016). “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”.

Xu, R. (2015). “Terminology Preparation for Simultaneous Interpreters”. PhD thesis, University of Leeds.

Yu, D. and Deng, L. (2015). *“Automatic Speech Recognition”*. A Deep Learning Approach, Springer-Verlag London, 2015.

Zahorian, S.A., Zimmer, A.M. and Meng, F. (2002). "Vowel Classification for Computer based Visual Feedback for Speech Training for the Hearing Impaired". In ICSLP 2002.

Zanettin, F. (2001). “Informatica e Traduzione”. In *Traduzione, revisione e localizzazione nel terzo millennio: da e verso l'inglese.* (eds.) C. Monacelli, Milan: Franco Angeli Editore: 19-45.

Zanettin, F. (2013). “Corpus methods for descriptive Translation Studies”. *Procedia—Social and Behavioral Sciences*, 95: 20-32.

Zetzsche, J. “The Age of Artificial Intelligence: Why translators are going to be the ones to turn off the lights in the offices after everyone else has long gone home”. *TeTra5 Conference*, Forli Campus, University of Bologna, 15th March 2019.

Other references

Colah’s Blog (1915). “Understanding LSTM Networks”. URL address: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. (Last visited on 15/12/2020).

CORDIS (Community Research and Development Information Service) of the European Commission. (2017). EU-BRIDGE. URL address: https://cordis.europa.eu/project/rcn/101838_en.html (Last visited on 15/12/2020).

DARPA-BOLT. Broad Operational Language Translation (BOLT), United States. URL address: https://www.darpa.mil/program/broad-operational-language-translation (Last visited on 15/12/2020).

DARPA-GALE. Global Autonomous Language Exploitation (GALE), United States. URL address: http://www.speech.sri.com/projects/GALE/ (Last visited on 15/12/2020).

DeepL (2020). DeepL Translate. Property of DeepL GmbH. URL address: https://www.deepl.com/translator

Descript (2020). Descript API and Web Interface. Property of Descript, 385 Grove St. San Francisco, CA 94102 (United States). URL address: https://www.descript.com/

ELRA (2015). European Language Resources Association. URL address: http://www.elra.info/en/projects/archived-projects/tc-star/. (Last visited on 15/12/2020)

EPTIC. European Parliament Translation and Interpreting Corpus. URL address: https://corpora.dipintra.it/eptic/?section=home (Last visited on 17/12/2020)

European Commission (2004-2007). TC-STAR. Technology and Corpora for Speech to Speech Translation. http://www.tcstar.org/

European Commission (2016). EU Audiovisual Media Services Directive. 25 May 2016. URL address: https://ec.europa.eu/digital-single-market/en/news/proposal-updated-audiovisual-media-services-directive (Last visited on 16/12/2020)

European Parliament channel on YouTube (2020). The official EP channel. URL address: https://www.youtube.com/user/EuropeanParliament

European Parliament of the European Union (2020). The official Web portal of the European Parliament. URL address: https://www.europarl.europa.eu/portal/en

European Parliament Plenary Sessions (2017-2020). The official channel of EPPS. URL address: https://www.europarl.europa.eu/plenary/en/home.html

FAO channel on YouTube (2020). The official FAO channel. URL address: https://www.youtube.com/user/FAOoftheUN

FAO TERM (2019-20). The FAO Terminology Portal. URL address: http://www.fao.org/faoterm/en/

Food and Agriculture Organization (FAO) of the United Nations (2020). The official Web portal of the Food and Agriculture Organization of the United Nations. URL address: http://www.fao.org/home/en/

Google Cloud Speech To Text (2020). *Google Speech Recognition and Speech to Text* technology. Property of Google, Inc. Gordon House, Barrow Street, Dublin 4, Ireland. URL address: https://cloud.google.com/speech-to-text (Last visited on 16/12/2020)

ILSA Project (2017-2020). The Interlingual Live Subtitling for Access Project. URL address: http://ka2-ilsa.webs.uvigo.es/ (Last visited on 16/12/2020)

International Organization for Standardization (2018). *ISO/IEC DIS 20071-23*. Information technology — User interface component accessibility — Part 23: Visual presentation of audio information (including captions and subtitles). URL address: https://www.iso.org/standard/70722.html (Last visited on 16/12/2020)

International Workshop on Spoken Language Translation (2013). URL address: http://www.iwslt2013.org.

JBI Studios' Blog (2020). "Subtitles Translation 101: Time-Stamping, Time-Coding & Spotting". URL address: https://jbilocalization.com/blog/subtitles-translation-101-time-stamping-time-coding-spotting/ (Last visited on 16/12/2020)

KantanMT Blog (2015). "What is Translation Error Rate (TER)?". URL address: https://kantanmtblog.com/2015/07/28/what-is-translation-error-rate-ter/ (Last visited on 15/12/2020).

SDL (2016). *SDL Research Survey 2016*: Translation Technology Insights. URL address: https://www.sdltrados.com/landing/lsp/Translation-Technology-Insight.html (Last visited on 15/12/2020).

Seventh Framework Programme of CORDIS. URL address: http://cordis.europa.eu/fp7/. (Last visited on 15/12/2020).

Speaker Hub (2017). "Your speech pace: guide to speeding and slowing down". Available at: https://medium.com/@speakerhubHQ/your-speech-pace-guide-to-speeding-and-slowing-down-be150dcb9cd7 (Last visited on 16/12/2020)

Towards Data Science (2019a). "Speech recognition is hard" (written by Shreya Amin). URL address: https://towardsdatascience.com/speech-recognition-is-hard-part-1-258e813b6eb7. (Last visited on 15/12/2020).

Towards Data Science (2019b). "Introduction to Hidden Markov Models. We present Markov Chains and the Hidden Markov Model." (written by Tomer Amit). URL address: https://towardsdatascience.com/introduction-to-hidden-markov-models-cd2c93e6b781. (Last visited on 15/12/2020).

Towards Data Science (2019c). "Neural Machine Translation. A guide to Neural Machine Translation using an Encoder Decoder structure with attention. Includes a detailed tutorial using PyTorch in Google Colaboratory" (written by Quinn Lanners). URL address: https://towardsdatascience.com/neural-machine-translation-15ecf6b0b. (Last Visited on 16/12/2020).

TranslateFX's Blog (2020). "What is Neural Machine Translation & How does it work?" (written by Sam Yip). URL address: https://www.translatefx.com/blog/what-is-neural-machine-translation-engine-how-does-it-work. (Last visited on 16/12/2020)

Ubiqus (2020). "Traduzione automatica neuronale (NMT). Che cos'è la traduzione automatica neuronale?". URL address: https://www.ubiqus.com/it/tecnologie/nmt-traduzione-automatica-neurale/#:~:text=La%20traduzione%20automatica%20neuronale%20o,partenza%20per%20alcune%20traduzioni%20professionali. (Last visited on 16/12/2020)

UKEssays (2018). "Three Circle Model of World Englishes English". Retrieved from https://www.ukessays.com/essays/english-literature/three-circle-model-of-world-englishes-english-literature-essay.php?vref=1 (Last visited on 16/12/2020)

UN Committee on Economic, Social and Cultural Rights (1999). General Comment 12. E/C.12/1999/5. (General Comments). URL address: https://www.escr-net.org/resources/general-comment-12#:~:text=The%20obligation%20to%20fulfil%20(facilitate,their%20livelihood%2C%20including%20food%20security.&text=This%20obligation%20also%20applies%20for,of%20natural%20or%20other%20disasters. (Last visited on 16/12/2020)

United Nations (1948). *Universal Declaration of Human Rights*. Paris. URL address: https://www.un.org/en/universal-declaration-human-rights/ (Last visited on 16/12/2020)

United Nations (2006). *UN Convention on the Rights of Persons with Disabilities*. URL address: https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html (Last visited on 16/12/2020)

United Nations (2014). *General Comment 2 on Article 9* released by UN Committee on the Rights of Persons with Disabilities. URL address: https://www.ohchr.org/en/hrbodies/crpd/pages/gc.aspx. (Last visited on 16/12/2020)

United Nations (2020). The official Web portal of the United Nations. URL address: https://www.un.org/en/

United Nations channel on YouTube (2020). The official UN channel. URL address: https://www.youtube.com/user/unitednations

Vocapia Research (2020). VoxSigma Speech to Text Software Suite. URL address: https://www.vocapia.com/voxsigma-speech-to-text.html (Last visited on 16/12/2020)

YobiYoba (2020). *YobiYoba Web service*. Property of Yobinext SAS, Parc Orsay Université, 91400 Orsay, France. URL address: https://www.yobiyoba.com/en/

Appendix A

DATABASE

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Label	**Language**	**Title**	**Native**	**Organization**	**Domain/Subdomain**	**Interference**	**Duration**	**Link URL**	**Year**	**Gender**	**Nationality**	**Speed**	**Pitch**	**Speaker**
EN_001_FAO_NN	English	FAO Director-General's speech at the Youth Employment in Agriculture Conference	No	FAO	Climate Change/Agriculture	Yes (applauses)	00:11:29	https://youtu.be/_J_DhMwnboU	2018	Male	Brasil	Slow	Medium	Jose Graziano da Silva
EN_002_FAO_NN	English	Director Parviz Koohafkan' Speech	No	FAO	Climate change/Agriculture	No	00:02:23	https://www.youtube.com/watch?v=eUQb89NkcEo	2016	Male	Iran	Slow	Medium	Parviz Koohafkan
EN_003_FAO_NN	English	Raja Devasish Roy, member of UNPFII sees national-level coordination with FAO as key	No	FAO	Climate Change/Agriculture	No	00:05:06	https://www.youtube.com/watch?v=TukVdlVr4w8	2015	Male	Bangladesh	Slow	Medium	Raja Devasish Roy
EN_004_FAO_NN	English	Statement by Director General FAO Jose Graziano da Silva, APRC 34, 2018	No	FAO	Climate change/Food Production	No	00:10:49	https://youtu.be/6Eeh-iKbHus	2018	Male	Brasil	Slow	Medium	Jose Graziano da Silva
EN_005_FAO_NN	English	Remarks by H.E Shafiul Alam, Land Ministry of Bangladesh	No	FAO	Climate chage/Agriculture	No	00:01:52	https://youtu.be/vsC3L8OuHT4	2015	Male	Bangladesh	Slow	Low	Shafiul Alam
EN_006_FAO_NN	English	Bharrat Jagdeo addresses the opening of the 5th World Forest Week	No	FAO	Climate change/Forestry	No	00:05:17	https://youtu.be/Sf0WasqI6kk	2016	Male	Ghana	Slow	Medium	Bharrat Jagdeo
EN_007_FAO_NN	English	Remarks by Indonesia's Minister for Marine Affairs and Fisheries	No	FAO	Climate change/Fishery	No	00:01:38	https://youtu.be/wcOPhvoYLN8	2017	Female	Indonesia	Average	Medium	Susi Pudjiastuti
EN_008_FAO_NN	English	Hina Rabbani Khar, Former Minister of Finance and former Minister of Foreign Affairs of Pakistan	No	FAO	Climate chage/Agriculture	No	00:01:22	https://youtu.be/13fz1jQ43fs	2018	Female	Pakistan	Fast	Medium	Hina Rabbani Khar
EN_009_FAO_NN	English	Global Soil Partnership interviews - Samuel Gameda	No	FAO	Climate chage/Soil Management	Yes (breaks with music)	00:06:23	https://youtu.be/jlwsFs2yI38	2013	Male	Ethiopia	Average	Low	Samuel Gameda
EN_010_FAO_NA	English	Global Oceans Action Summit -- Feedback on The Economist World Ocean Summit	Yes	FAO	Climate change/Oceans	No	00:14:27	https://youtu.be/hNcLocj_I7c	2014	Male	UK/Hong Kong	Average	Medium	Charles Goddard
EN_011_FAO_NN	English	Remarks by Moses Vilakati, Minister for Agriculture of Swaziland	No	FAO	Climate change/Fishery/Forestry	No	00:04:20	https://youtu.be/KZ9iiWTde98	2015	Male	Swaziland	Average	Medium	Moses Vilakati
EN_012_FAO_NA	English	H.E. Roger Clarke (Jamaica)	Yes	FAO	Climate change/Nutrition	No	00:01:55	https://youtu.be/-5FFqtX8OA4	2013	Male	Jamaica	Average	Low	Roger Clarke
EN_013_FAO_NA	English	Pretoria Symposium 2015 opening speech by M. Burke (ICAR)	Yes	FAO	Climate change/Farming	No	00:04:07	https://youtu.be/FQ3vIeApKJY	2015	Male	South Africa	Average	Medium	Martin Burke
EN_014_FAO_NN	English	Remarks by Liberia's Minister for Agriculture, Florence Chenoweth, at FAO.	Yes	FAO	Climate Change/Agriculture	No	00:05:17	https://youtu.be/pLeZNX7Aois	2015	Female	Liberia	Slow	Low	Florence Chenoweth
EN_015_FAO_NN	English	Global Oceans Action Summit - Árni M. Mathiesen, Assistant Director-General, FAO Fisheries	No	FAO	Climate change/Oceans Management	No	00:10:36	https://youtu.be/doBVO5_5gtU	2014	Male	Iceland	Average	Medium	Árni M. Mathiesen
EN_016_FAO_NA	English	World Forest Week special event: Closing remarks by Jeffrey Y Campbell, Manager, FFF	Yes	FAO	Climate change/Forestry	Yes (microphone/other speaker first seconds)	00:06:45	https://www.youtube.com/watch?v=kc-KGbjAWLE	2018	Male	US	Average	Medium	Jeffrey Y. Campbell
EN_017_FAO_NA	English	Remarks by Ireland's Minister of State for Food, Forestry and Horticulture	Yes	FAO	Climate change/Plant Health	No	00:06:25	https://youtu.be/bvGD0xXjmX8	2017	Male	Ireland	Average	Low	Andrew Doyle
EN_018_FAO_NN	English	Hoesung Lee addresses the opening of the 5th World Forest Week	No	FAO	Climate change/Forestry	No	00:10:05	https://youtu.be/drICefggrQw	2017	Male	South Korea	Average	Medium	Hoesung Lee
EN_019_FAO_NN	English	Dr. Braulio Dias, CBD Executive Secretary - International Green Week, Berlin 2014	No	FAO	Food Security/Biodiversity	No	00:05:33	https://youtu.be/sakS7zi68Fs	2015	Male	Brasil	Average	Medium	Braulio Dias
EN_020_FAO_NA	English	Remarks by H.R.M. King Letsie III of the Kingdom of Lesotho, FAO Special Ambassador for Nutrition	Yes	FAO	Food Security	Yes (Road traffic)	00:04:04	https://youtu.be/nwH5h6PwURU	2018	Male	Lesotho	Slow	Medium	King Letsie III
EN_021_FAO_NN	English	H.E. Vidar Helgesen addresses the opening of the 5th World Forest Week	No	FAO	Climate Change/Forestry	No	00:08:43	https://youtu.be/TRMx8wLU4cA	2016	Male	Norway	Average	Medium	Vidar Helgesen
EN_022_FAO_NN	English	Sławomir Mazurek addresses the opening of the 6th World Forest Week	No	FAO	Climate Change/Forestry	No	00:05:32	https://youtu.be/l_rZtfuk9cs	2018	Male	Poland	Slow	Medium	Sławomir Mazurek
EN_023_FAO_NN	English	David Kaatrud gives WFP statement on Zero Hunger to FAO APRC 34, 2018	No	FAO	Food Security	No	00:13:58	https://youtu.be/Lo_LpLQ3jFI	2018	Male	Belgium	Average	Medium	David Kaatrud
EN_024_FAO_NN	English	FAO-Nobel Peace Laureates Alliance - Kofi Annan	No	FAO	Food Security	No	00:02:24	https://youtu.be/kz6RF9ZN0HQ	2016	Male	Ghana	Slow	Medium	Kofi Annan
EN_025_FAO_NN	English	Director-General’s remarks at the Regional Symposium on Agroecology in Europe and Central Asia	No	FAO	Agriecology/Climate Change	Yes (applauses)	00:07:52	https://youtu.be/12kO0vOLomY	2017	Male	Brasil	Slow	Medium	Jose Graziano da Silva
EN_026_FAO_NN	English	HRH Prince Laurent of Belgium addresses the opening of the 5th World Forest Week	No	FAO	Forestry/Climate Change	Yes (coughing)	00:14:38	https://youtu.be/eFrTPr87Uuw	2016	Male	Belgium	Slow	Medium	Prince Laurent of Belgium
EN_027_FAO_NA	English	Dan Gustafson, FAO Deputy-Director Address to the Global Agriculture and Food Security Program	Yes	FAO	Food Security	No	00:05:23	https://youtu.be/gBS86AK949g	2017	Male	US	Average	Medium	Dan Gustafson
EN_028_FAO_NN	English	Remarks by T.H. Bernhard Esau, Minister of the Republic of Namibia	Yes	FAO	Fishery	No	00:06:46	https://youtu.be/3UL-s2p31Y8	2016	Male	Namibia	Average	Medium	Bernhard Esau
EN_029_FAO_NA	English	How the use of plant genetic resources helped India to fight hunger -- K.C. Bansal	Yes	FAO	Food Security	No	00:06:17	https://youtu.be/4TcP8Um98q0	2014	Male	India	Fast	Medium	K.C. Bansal
EN_030_FAO_NN	English	Colombo’s city region food system: The challenges, current situation and way forward	Yes	FAO	Food Security	No	00:06:54	https://youtu.be/bGn5PQRgbTw	2015	Male	Sri Lanka	Average	Medium	Ruwan Wijayamuni
EN_031_FAO_NA	English	Fall Armyworm Monitoring and Early Warning System (FAMEWS) COAG 26 SPEAKER'S CORNER	Yes	FAO	Pest Control/Agriculture	Yes (echo)	00:03:48	https://youtu.be/wsQrbQs_32I	2018	Male	US	Fast	Medium	Allan Hruska
EN_032_FAO_NN	English	Nauru Country Statement, FAO APRC 34, 2018	Yes	FAO	Food Security/Climate change		00:06:27	https://youtu.be/2HJoWizad3E	2018	Male	Republic of Nauru	Average	Medium	Lionel Rouwen Aingimea
EN_033_FAO_NA	English	Joan Burton, Deputy Prime Minister of Ireland	Yes	FAO	Food Security	Yes (background noise)	00:01:27	https://youtu.be/8Z8A3MA5DA4	2016	Female	Ireland	Average	Medium	Joan Burton
EN_034_FAO_NA	English	Australia Country Statement to FAO APRC 34, 2018	Yes	FAO	Food Security	No	00:10:27	https://youtu.be/_gEZXJwgfVg	2018	Male	Australia	Average	Medium	Matthew Worrell
EN_035_EP_NN	English	Greta Thunberg's emotional speech to EU leaders	No	EURP	Climate Change	Yes (tears) and applauses	00:04:11	https://www.youtube.com/watch?v=FWsM9-_zrKo	2019	Female	Sweden	Average	Low	Greta Thunberg
EN_036_EP_NA	English	EU Hypocrisy on Climate Change	Yes	EURP	Climate Change	No	00:01:17	https://www.youtube.com/watch?v=u2-Nv9Awgak	2009	Male	UK	Average	Medium	Daniel Hannan
EN_037_EP_NN	English	Debate of 18 Nov 2018 at EP	No	EURP	Climate Change	No	00:04:41	https://www.youtube.com/watch?v=Xucy-N-CBlA	2018	Male	Slovakia	Average	Medium	Maros Sefcovic
EN_038_EP_NN	English	Debate of 18 Nov 2018 at EP	No	EURP	Climate change	No	00:05:22	https://www.youtube.com/watch?v=Xucy-N-CBlA	2018	Male	Spain	Fast	Medium	Miguel Arias Canete
EN_039_EP_NN	English	Speech by MACanete Debate of 13 March 2019 Part1	No	EURP	Climate change	No	00:07:02	https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20190313&detailBy=date	2019	Male	Spain	Fast	Medium	Miguel Arias Canete
EN_040_EP_NN	English	Speech by MGCiot Debate of 13 March 2019 Part1	No	EURP	Climate change	No	00:08:14	https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20190313&detailBy=date	2019	Famale	Romania	Average	Medium	Maria Gabriela Ciot
EN_041_EP_NN	English	Speech by MGCiot Debate of 13 March 2019 Part2	No	EURP	Climate change	No	00:01:11	https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20190313&detailBy=date	2019	Female	Romania	Average	Medium	Maria Gabriela Ciot
EN_042_EP_NN	English	Speech by MACanete Debate of 13 March 2019 Part2	No	EURP	Climate change	No	00:01:57	https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20190313&detailBy=date	2019	Male	Spain	Average	Medium	Miguel Arias Canete
EN_043_EP_NN	English	Speech by BEickout Debate of 13 March 2019	No	EURP	Climate Change	Yes (applauses)	00:04:07	https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20190313&detailBy=date	2019	Male	The Netherlands	Fast	High	Bas Eickhout
EN_044_EP_NN	English	Speech by GJGerbrandy Debate of 13 March 2019	No	EURP	Climate change	No	00:02:19	https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20190313&detailBy=date	2019	Male	The Netherlands	Average	Medium	Gerben-Jan Gerbrandy
EN_045_EP_NN	English	Speech by UBullmann Debate of 13 March 2019	No	EURP	Climate change	Yes (applauses)	00:03:26	https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20190313&detailBy=date	2019	Male	Germany	Average	High	Udo Bullmann
EN_046_EP_NA	English	Speech by LBoylan Debate of 13 March 2019	Yes	EURP	Climate change	No	00:01:32	https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20190313&detailBy=date	2019	Female	Ireland	Average	High	Lynn Boylan
EN_047_FAO_NA	English	Pacific Islands Forum (PIF) Secretary General on climate change	Yes	FAO	Climate change	No	00:01:20	https://www.youtube.com/watch?v=ExfyZodJn9k	2017	Female	Australia/PNG	Average	Low	Meg Taylor
EN_048_UKG_NA	English	Prime Minister's Speech on the Environment	Yes	UK Government	Climate Change	Yes (applauses)	00:25:23	https://www.youtube.com/watch?v=yguGDwVTtE4	2018	Female	UK	Fast	Medium	Theresa May
EN_049_UKP_NA	English	Jeremy Corbyn’s Call for Climate Emergency which was endorsed by the UK parliament on 1:st of May	Yes	UK Parliament	Climate change	Yes (noise)	00:14:22	https://www.youtube.com/watch?v=wA3N1Nq0k1I	2019	Male	UK	Average	Medium	Jeremy Corbyn
EN_050_UN_NA	English	President Barack Obama at UN Climate Change Summit	Yes	UN	Climate change	No	00:10:40	https://www.youtube.com/watch?v=WYga2qRnY2w	2009	Male	US	Average	Medium	Barak Obama
EN_051_FAO_NA	English	Fall Armyworm Monitoring and Early Warning System (FAMEWS) COAG 26 SPEAKER'S CORNER	Yes	FAO	Climate change/Pest Control	Yes (echo)	00:12:00	https://youtu.be/wsQrbQs_32I	2018	Male	US	Fast	Medium	Keith Cressman
EN_052_UN_NN	English	UN Chief on Climate Change and his vision for the 2019 Climate Change Summit	No	United Nations	Climate change	No	00:07:35	https://www.youtube.com/watch?v=Jsi5Vp_6tdE	2018	Male	Portugal	Average	Medium	António Guterres
EN_053_UN_NN	English	UN Chief on Climate Change and his vision for the 2019 Climate Change Summit	No	United Nations	Climate change	No	00:03:46	https://www.youtube.com/watch?v=Jsi5Vp_6tdE	2018	Female	Sri Lanka	Fast	Medium	Jayathma Wickramanayake
EN_054_UN_NA	English	Charles: Humanity faces no greater threat than climate change	Yes	United Nations	Climate change	No	00:01:12	https://www.youtube.com/watch?v=P0ifN4K8FTQ	2015	Male	UK	Average	Medium	Prince Charles of United Kingdom
EN_055_UN_NA	English	Leonardo DiCaprio Delivers Powerful Climate Change Speech At The UN	Yes	United Nations	Climate change	Yes (applauses)	00:11:28	https://www.youtube.com/watch?v=qkZ13cVUbJs&t=15s	2016	Male	US	Average	Medium	Leonardo Di Caprio

Appendix B

Annotation Instructions – Inter-Annotator Agreement

BEFORE YOU START

The following instructions for annotations are aimed at defining a protocol for the insertion and validation of annotations by part of a pool of external annotators to this study. These instructions are intended to provide for sufficiently clear and simple guidelines for the insertion of annotations in an Excel file of comparing. The file includes the reference transcription for the video in question and the automatic speech transcription generated by the SR software: namely, VoxSigma by Vocapia and Google Speech Recognition engine via YouTube or Descript interface. Both videos are public speeches at a conference on Climate Change held by an official at FAO of the United States (the speaker may be native or non-native). Please note that, for convenience, the segmentation of automatic software transcriptions will follow the segmentation generated by VoxSigma. The final objective of this phase will be to validate the protocol followed in this study to annotate errors by automatic Speech Recognition software, including the taxonomy of errors defined.

In the two **Compare Excel files** provided to you, from the left to the right, you will find:

- **The Segment ID:** the number of segment
- **The Time Stamp:** the time segmentation provided by the software
- **The Gold Standard:** the correct, reference transcription
- **VoxSigma or YouTube Transcription:** the transcription generated by the SR software
- **Gross-Grained Error:** it is the macro category for the taxonomy of errors (i.e., Insertion, Deletion or Substitution)
- **Fine-Grained Error:** it includes the fine-grained subcategories of errors according to the classification indicated below.
- **Notes:** it includes the notes or comments containing further details for defining the type of errors (to be edited by the annotator).
- **Error Seriousness:** it includes the definition of errors severity as "Serious" or as "Not serious".
- **HIT Number:** this column indicates the "hit" number, *i.e.* whether the automatic segment transcription is totally matching the reference segment (value "1") or not ("0").

ANNOTATION INSTRUCTIONS

After reading the introduction above, you are now ready to **fill in the Columns from E to I** as per the instructions below.

Gross-Grained Error (Column E)

Please indicate here if the macro errors category is a **Deletion**, **Substitution** or **Insertion** type of error. For convenience, the software's omitted or inserted words or expressions are already given in the Column G, and they are reported between square brackets. In particular:

1. Enter **Deletion** in the cell when the software transcription has omitted one word or an entire expression. In addition to this, report the deleted word or expression in red in Column E between square brackets by following the Gold Standard for reference:

e.g.: [And] I think we had very positive work [for doing] during those 2 days here (reference text in Column E was: *"And I think that we had very productive work for doing during those two days here"*)

Attention: when two occurrences of errors show up in the same segment, the segment is to be repeated in the line below, considering them as two errors. In the example above, the errors are two so it is necessary to generate two different segment lines in the Excel file (even when the error type is the same).

2. Enter **Substitution** when the software transcription has replaced one word or an entire expression with another word or expression. In addition to this, indicate the deleted word or expression in red in Column E by following the Gold Standard for reference:

e.g.: allowed me to start break into protocols (r0eference text in Column E was: *"Allow me to start breaking the protocols"*).

3. Enter **Insertion** when the software transcription has added or inserted one word or an entire expression with respect to the gold standard. In addition to this, indicate the inserted word or expression in red in Column E by following the Gold Standard for reference:

e.g.: Taking into account of the main concern of the issue of migration (reference text in Column E was: *"taking into account the main concern of the issue of migration"*).

Fine-Grained Error (Column F)

Under this column indicate the fine-grained error category as per the classification below: Speech-related features (here called **Disfluency**), **Grammar**, **Lexis**, **Terminology**, **Prosody**. Examples:

Disfluency: please indicate errors under this subcategory when the error is any of various breaks, irregularities, or non-lexical vocables which occur within the flow of otherwise fluent speech. These include "false starts", i.e. words and sentences that are cut off mid-utterance; phrases that are restarted or repeated and repeated syllables; "fillers", i.e. grunts or non-lexical utterances such as *"huh"*, *"uh"*, *"erm"*, *"um"*, *"well"*, *"so"*, *"like"*, *"you know"*, and *"hmm"*.

e.g.: "So [well] but what we can do, we, FAO…" (Example of Deletion, Disfluency)

e.g.: "…but [but] in order to tackle with drought…" (Example of Deletion, Disfluency)

Grammar: indicate errors under this subcategory when the error relates to the language set of rules or syntax rules. For example: verb tense (define/defined), pronouns (this/these), prepositions, etc. For convenience errors related to lowercase/uppercase letters and punctuation are not taken into account.

e.g.: "FAO develop an initiative aimed at building up…" (in reference text, you have "developed", so this is an example of Substitution, Grammar error).

Lexis: under this subcategory, please enter errors relating to a wrong recognition of common lexis or vocabulary. Errors in numbers transcription are also account for in this subcategory: for example, *"fourty/fourteen"*.

e.g.: "And I will be cheering the 23rd session of conference on… (in reference text: "and I will be chairing the twenty-third session of conference on…"; this is example of Substitution, Lexis error).

Terminology: under this subcategory are the errors relating to a wrong recognition of specific vocabulary, proper names, specific terminology, names of committees, research groups, organizations, international/regional initiatives.

> e.g.: "Foul participated in the initiative in 2019…" (in reference text, you have "FAO participated in the initiative in 2019…"; example of Substitution, Terminology error).
>
> e.g.: "I am pleased to open colorful (in reference text: "I am pleased to open COFO"; in this case we have an example of Substitution – Terminology error).

Prosody: under this subcategory are the errors relating to intonation or stress. As punctuation is not considered in the annotation process and in the counting of errors, the only type of error pertaining to prosody is the absence/addition of a question mark (*"?"*) when it is available/not available in the reference transcription text.

> e.g.: "And I asked Geraldine can FAO participate[?] and she said yes" (in reference text, we have: "And I asked Geraldine, can FAO participate? And she said Yes).

Notes (Column G)

Please include here further details, if relevant to the analysis. For example, in case of errors like "this/these", it is possible to specify that this is an example of minimal pair or near-homophone.

Error Seriousness (Column H)

Specify whether the error is **"Serious"** or **"Not Serious"** to your judgment. Please note that serious errors are recognition mistakes that do not allow the general understanding of the segment unit in question.

Examples of serious errors: *"fourteen"* instead of *"fourty"*; *"foul"* instead of *"FAO"* etc.

Examples of not serious errors: *"this"* instead of *"these"* or *"develop"* instead of *"develops"*, etc.

Hit Number (Column I)

Enter the value "1" when the automatic segment transcription is totally matching the reference segment or the value "0" when not.

If required, you can consult or watch the video on File 1, by visiting the URL:

https://www.youtube.com/watch?v=eUQb89NkcEo

For the video on File 2, go to URL: https://youtu.be/-5FFqtX8OA4

THANK YOU FOR YOUR CONTRIBUTION!

Video 002

SEGMENT ID	TIMESTAMP	REFERENCE TRANSCRIPTION	SR TRANSCRIPT	COARSE-GRAINED ERROR	FINE-GRAINED ERROR	ERROR SERIOUSNESS	NOTES
1	00:00:04,200 --> 00:00:08,130	The globally important Agricultural Heritage System Initiative	the globally important agricultural heritage system initiative	Substitution; Deletion; Insertion	Grammar, Disfluency, Lexis, Terminology, Prosody	Serious, Not Serious	
2	00:00:09,900 --> 00:00:19,270	is about farmers who are in remote areas, and they have created through centuries very	is about farmers who are in remote areas and they have created through centuries very				
3	00:00:19,270 --> 00:00:24,310	very outstanding agricultural systems to	very outstanding agriculture systems to				
4	00:00:25,220 --> 00:00:30,770	get their food security and livelihood from the these systems.	get their food security and livelihood from the these systems				
5	00:00:32,189 --> 00:00:38,490	Ehm, we have identified some two hundred systems around the world which are unique in different	we have identified some 200 systems around the world which are unique in different				
6	00:00:39,040 --> 00:00:45,530	aspects of food security, biodiversity, indigenous knowledge, cultural diversity,	aspects food security biodiversity indigenous knowledge cultural diversity				
7	00:00:46,000 --> 00:00:51,269	and of course landscape diversity. Ehm, for implementing this program.	and of course landscape diversity for implementing this program				
8	00:00:51,590 --> 00:00:53,970	We have been working at three levels,	we have been working at three levels				
9	00:00:54,110 --> 00:00:59,260	at global level, to get the recognition of this agricultural Heritage, similar to	at global level to get the recognition of these agricultural heritage similar to				
10	00:00:59,260 --> 00:01:07,510	World Heritage Sites of UNESCO and at national level, to review national policies	World Heritage Sites of UNESCO and at national level to review national policies				
11	00:01:07,510 --> 00:01:11,190	in food security, indigenous people, to enable us	in food security indigenous people to enable us				
12	00:01:12,070 --> 00:01:19,970	to actually help better these marginal farms and poor farmers and at local level,	to actually help better these marginal farms and poor farmers and at local level				
13	00:01:19,970 --> 00:01:23,190	particularly looking at goods and services	particularly looking at goods and services				
14	00:01:23,289 --> 00:01:28,479	these farmers are providing to humanity, while maintaining natural resources	these farmers are providing to humanity by maintaining natural resources				
15	00:01:29,200 --> 00:01:36,950	and, of course, managing biodiversity, in particular by true eco-labelling, to	and of course managing by diversity in particular by true eco labeling to				
16	00:01:37,950 --> 00:01:43,110	sustainable tourism through enhancement of their productivity, and of course	sustainable tourism through enhancement of their productivity and of course				
17	00:01:43,110 --> 00:01:45,690	Community Development around the world.	community development around the world				
18	00:01:46,130 --> 00:01:54,800	In Peru, we have had very much successes in the sense that Andean region of Peru is	in Peru we have had very much successes in the sense that andean region of Peru is				
19	00:01:54,960 --> 00:02:00,730	populated with many many different tribes and groups and they have been developing	populated with many many different tribes and groups and they have been developing				
20	00:02:00,910 --> 00:02:06,840	fascinating agricultural systems which are unique in the world and they have many	fascinating agricultural systems which are unique in the world and they have many				
21	00:02:06,840 --> 00:02:12,510	many crops and products which actually have made to the market	many crops on products which actually have made to the market				
22	00:02:13,894 --> 00:02:17,725	to the international market, by labeling and by promotion.	to the international market by labeling and promotion				

Video 012

SEGMENT ID	TIMESTAMP	REFERENCE TRANSCRIPTION	SR TRANSCRIPT	COARSE-GRAINED ERROR	FINE-GRAINED ERROR	ERROR SERIOUSNESS	NOTES
1	00:00:07,320 --> 00:00:11,350	A major problem we face has to do with the fact that	A major problem we face as they do with the fact that	Substitution, Deletion, Insertion	Grammar, Disfluency, Lexis, Terminology, Prosody	Serious, Not Serious	
2	00:00:12,420 --> 00:00:14,330	as far as family farming goes, uhm	as far as family farming goes [uhm]				
3	00:00:15,510 --> 00:00:20,640	the major production is in that general area	the major production is in that generally aware				
4	00:00:20,760 --> 00:00:23,690	where our small farmers don't have the wherewithal	I was small farmers don't have the wherewithal				
5	00:00:24,660 --> 00:00:27,840	to really produce in an efficient	to really produce in an efficient				
6	00:00:28,320 --> 00:00:31,430	way. They have marketing problems	way. They have marketing problems				
7	00:00:31,890 --> 00:00:36,450	and all that. And when you, when one looks at	and all that. And when you when one looks at				
8	00:00:37,320 --> 00:00:40,950	the diet, it's more skewed towards uhm	the debt. It's more skewed towards				
9	00:00:42,570 --> 00:00:46,110	staples like starches and so forth.	staples like statues and sort of it				
	00:00:42,570 --> 00:00:46,110	staples like starches and so forth.	staples like statues and sort of it				
10	00:00:46,920 --> 00:00:48,750	They, we have not yet been able to	that we have not yet been able to				
11	00:00:50,220 --> 00:00:55,200	let them afford a substantial amount of protein.	let them up afford a substantial amount of protein.				
12	00:00:55,200 --> 00:00:57,720	That is a major challenge for them.	That is my major challenge for them.				
13	00:00:57,720 --> 00:01:00,960	We are working on that as we speak.	We are working on that as we speak.				
14	00:01:00,960 --> 00:01:06,120	What we have done, we have been concentrating on small ruminants.	What we have done, we have been concentrating [on] the Noah smaller woman.				
15	00:01:06,120 --> 00:01:09,510	We are almost now self-sufficient with pork	We are almost no self-sufficient with pork				
16	00:01:11,040 --> 00:01:15,600	and we are working towards developing our dairy industry.	and we have working towards developing of the dairy industry.				
17	00:01:15,600 --> 00:01:18,880	So those that are here will help to	So those that year, will help to				
18	00:01:20,190 --> 00:01:25,260	improve nutrition. Away from that, we have put in	improve nutrition away from that we have put in				
19	00:01:25,380 --> 00:01:30,420	place legislation which will allow us as a	place legislation, which will allow us as a				
20	00:01:30,530 --> 00:01:34,260	government to dictate terms as to what is produced	government to dictate terms as to what is produced				
21	00:01:35,610 --> 00:01:39,090	for our people. We want to make sure that our people have safe	for our people we want to make sure that our people have safe				
22	00:01:39,090 --> 00:01:43,720	food and affordable food and we don't want them to have to	food and affordable food and we don't want them to have to				
23	00:01:44,580 --> 00:01:46,790	be in a position that they can't	be in a position that they can't				
24	00:01:47,310 --> 00:01:50,970	deal with shocks when you know, like when hurricanes	deal with sharks when you know like when hurricanes				
25	00:01:51,510 --> 00:01:54,900	and that sort of things happens. Okay, great.	and that sort of thing happens. Okay, great.				