ALESSANDRO GREGORI

Automatic Speech Recognition (ASR) for Communications with the Elderly and Non-hearing Users at Public Spaces: Speech-to-Text Technology for Live Subtitling and Accessibility

1. Introduction

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 universal right stemmed from the Universal Declaration of Human Rights (UDHR) of the United Nations (Paris, 1948) where the concepts of human dignity and access were established for the first time. According to the UDHR, the concept of access establishes the right to access to the essential resources required for a minimum standard of life quality. The right of accessibility was recently 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 which was 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. 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 provision 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 speech translation at an interinstitutional and inter-governmental level (Maslias 2017), not only in the EU, but also globally. Given the frequent, non-availability of interpreters or re-speakers for non-hearing people at the institutional level for any language combination, the application of ASR technology, combined with NMT, may possibly help in breaking down the barriers of communication within the global institutional context, where multilingualism certainly represents a fundamental pillar of Institutional Translation (Jopek Bosiacka 2013).

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 with the elderly (Goman 2017). Thanks to Speech to Text (STT) technology (and the production of live subtitles), it is possible to guarantee content accessibility for non-hearing old people at institutionally held conferences or speeches.

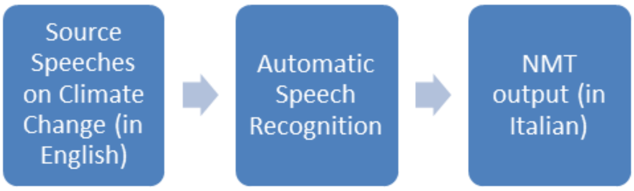
Source Speeches on Climate Change (in English)
Automatic Speech Recognition
NMT output (in Italian)

Figure 1. The research project's scenario.

While being part of a larger research project also involving the application of NMT and the analysis of subtitles in Italian, this study analyses official speeches hosted at international organisations on climate change and its effects on agricultural production. More specifically, the main research questions of this study are formulated as follows: (1) RQ1: 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?

To the best of my knowledge, this kind of speech has not been investigated so far in literature as a form of input data for ASR. The analysed data are collected in a multimedia database of audio/video materials and their relevant transcriptions into English as shown in Figure 1 above. Additionally, if in previous studies and projects in literature (ILSA project, TC-STAR, EU-Bridge, etc.), attention was mainly paid to the usage of ASR technology in combination with the intervention of either a subtitle editor or re-speaker, or an interpreter, in this study the human mediation role is eliminated by attempting to define a protocol for the usage of an entire ASR+NMT pipeline as shown in Figure 2 below.

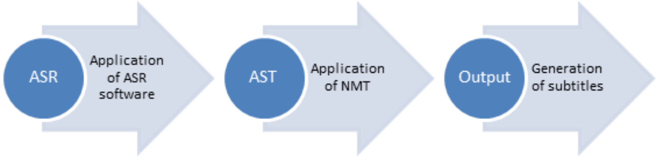

Figure 2. ASR-NMT-based pipeline methodology in this study.

2. Methods

In the methodology of this study, the construction of a database of files was the first step. For describing it in general terms, it is possible to make a few preliminary considerations with McEnery and Hardie (2012):

> Corpora may encode language produced in any mode--for example, there are corpora of spoken language and there are corpora of written language. In addition, some video corpora record paralinguistic features such as gesture ..., and corpora of sign language have been constructed. (McEnery and Hardie 2012, p. 3)

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. It is not therefore possible to use a definition of corpus linguistics given the specific nature of the database examined here. More specifically, the 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 consistent selection criteria, including the authenticity of texts (all speeches were authentic public discourse) and their representativeness. 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'. Representativeness is here guaranteed also in terms of Native/Non-Native speaker distribution. In addition to these requisites, other criteria were identified: a comparable institutional setting (hosting institutions are international organisations); topic and timespan (consistency is maintained in terms of topic and timespan: period 2013–2019); single speaker (mono-speaker-based and cover a similar institutional function/role); quality (all parts of the speech are clearly audible, with no interruptions and in optimal audio condition).

All fifty five audio/video files are official speeches on climate change given at the Food and Agriculture Organisation (F.A.O.)[1], the European Parliament[2] or the United Nations[3]. In particular, files are made publicly available on their official Websites or official channels for anyone willing to listen to or watch them. All these multimedia contents are therefore free and do not require any authorisation. The corpus of audio/video texts amounts to a total of 44,838 words[4] and a total duration of five hours, fifty three minutes and thirty four seconds[5]. The average length of each video/audio file is of six minutes and twenty six seconds. The speakers are fifty and they come from a total of thirty four countries. If the speech distribution is analysed further (though not relevant to the main RQs above), it is possible to observe that the gender composition is as follows: forty five speeches are held by male speakers and ten by female speakers. In this respect, it is evident 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 organisations generally sees a prevalence of male individuals (ISPI 2012). However, it should be remarked that, for the purposes of the data analysis, the gender representation is not relevant to the discussion of results and it is here given only for a better description of the database. At this stage, if the database is subdivided according to the Native/Non-Native categorisation, it is possible to see that the distribution of the speaker population per minutes of speech is as follows in Table 1.

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

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

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

[4] 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.

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

Group	Minutes	Percentage
Native	02:49:43	48%
Non Native	03:03:51	52%

Table 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. 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 (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, a 20% of the speech sample with a Fast speed rate (over 150 wpm), as shown in Figure 3 below.

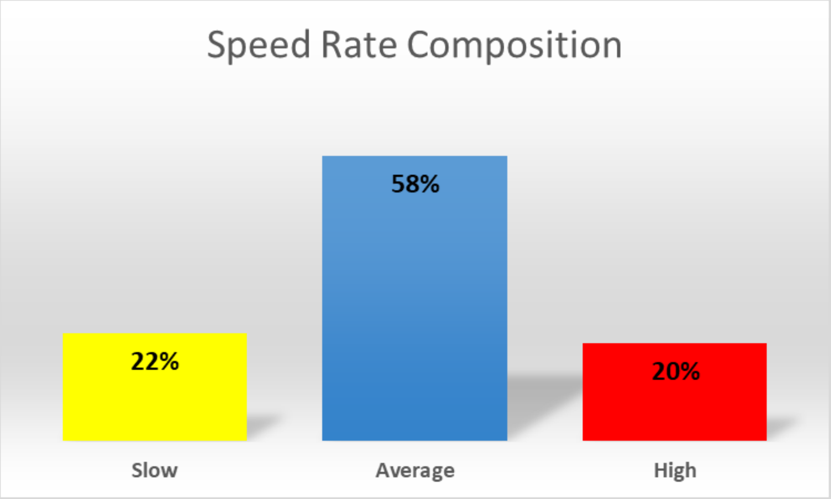

Figure 3. Composition of the speaker population according to the speed rate (wpm).

With reference to the ASR technology selection, the ASR solutions implemented had to comply with specific requirements. Most advanced software solutions available in the market today can better cope with the criticalities of speech recognition if compared to the past's technologies, and, in particular, 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). 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 (Saon and Chien 2012: 1–2). Other important features are the Cloud-based technology, minimum computer requirements, trainable functionality (i.e. the software can learn from previous data processing operations), as well as a fair price-capability ratio. Among various options, the selection was oriented towards VoxSigma suite developed by Vocapia Research, and Google Speech Recognition (GSR) engine (to be used via YouTube and Descript applications) provided that these ASR solutions responded to the requisites described above and for offering a good ratio between their cost and the capabilities offered. In this respect, it should also 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.

After selecting the most suitable ASR, a general protocol for the data processing workflow was established as shown in Figure 4 below.

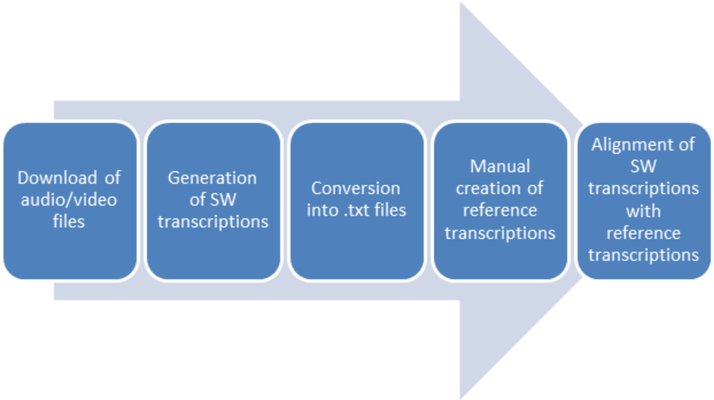
Download of audio/video files
Generation of SW transcriptions
Conversion into .txt files
Manual creation of reference transcriptions
Alignment of SW transcriptions with reference transcriptions

Figure 4. General workflow for data processing.

The workflow was organized into five steps, which can be described as follows: 1) Download of audio/video files on the PC locally in the *.avi* or *.mp4* format; 2) Generation of SW transcriptions by using DownSub[6] or the Descript app[7]**,** allowing executing the automatic transcription of files through GSR engine; 3) Conversion into .txt files (with no tags) by using the Export command in VoxSigma and Descript apps or SubtitleEdit.exe[8]; 4) Creation of the reference transcriptions (gold standard); 5) Alignment of SW transcriptions with reference transcriptions by following the time stamp organisation generated by VoxSigma. The alignment of texts was produced in Excel spreadsheets.

Regarding the transcription procedure followed, it should be commented that, for convenience, the manual reference transcriptions of all speeches were carried out starting from the automatic transcriptions generated by the software VoxSigma rather than making a transcription from scratch. This offered a valuable basis for quickly creating the final reference transcriptions, since it speeded up the process of

[6] https://downsub.com/
[7] https://www.descript.com/
[8] https://www.nikse.dk/subtitleedit

manual transcription. It should also be mentioned that 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, Russo et al. 2012: 57). In particular, in the present study, in order to avoid unnecessary complexities and to prevent transcription from being excessively time-consuming, it was decided to produce basic reference transcriptions. For an overview of the transcription conventions adopted in the present study, Table 2 below provides with further details.

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 2. Transcription conventions adopted in this study.

More specifically, the 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. 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). As far as the spelling convention is concerned, the study's gold standard transcriptions mostly followed the Interinstitutional Style Guide (European Union, 2020). Additionally, the uttered abbreviations for proper names, institutions, organisations or official programmes/initiatives used by the focused-on international organisations were 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.

Regarding the annotation of errors in ASR transcriptions of speeches, it was necessary to define an appropriate taxonomy of errors in order to properly analyse the output generated by ASR. The taxonomy was subdivided into two layers: Coarse-Grained Errors (Layer 1) and Fine-Grained Errors (Layer 2). First of all, it should be pointed out that, for its construction, 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'.

Layer 1 taxonomy identifies three main error typologies on an as much as possible objective way, by applying the literature most popular classification of errors based on the WER model. Within 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 the Table 3 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. Layer 1 for Taxonomy of Errors.

Layer 1 (Coarse-Grained Errors) should be considered as the main taxonomy 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 Inter-Annotator Agreement).

Layer 2 taxonomy is based on a fine-grained classification of errors built upon five main categories: Disfluency, Grammar, Lexis, Terminology, and Prosody. Before describing the set of rules used here to identify and classify the error types into five categories, it is essential to maintain that these categories are not intended to be objective nor 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, with the substitution of the adjective ‘their’ with ‘them’, the ambiguity between Grammar and Lexis does emerge. These error categories were mainly used for descriptive purposes in the study and were not aimed at evaluating accuracy in statistical terms. In synthesis, 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. Grammar 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’. Lexis includes all errors relating to lexical parts of the speech, thus including nouns, terms and also adjectives, as well as numbers and figures. The Terminology category accounts for specialized terminology errors relating to names of institutions, international initiatives, domain-specific terminology and also proper names. Finally, Prosody takes into consideration errors connected with intonation, i.e. with end-of-sentence question marks or exclamation marks.

Although useful for assessing ASR output, the Word Error Rate (WER) model is certainly 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. In particular, WER 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 5. 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. 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 probably 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 acronym three letters stand for *Number* (of words), *Edition* errors and *Recognition* errors. The overall score is calculated as follows:

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

CE (correct editions):

Assessment:

Figure 6. 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. For convenience, this study 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. 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. A certain degree of subjectivity is certainly associated to the process of annotation but, as defined below, the taxonomic scheme and error grading system implemented here was validated by means of an Inter-Annotator Agreement test. For examples of Serious or Not Serious errors, see Table 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 4. Error grading system.

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 calculations, respectively. This should help in better representing the severity differentiation of errors and in responding more efficaciously to the various applications of live subtitling. Generally, accuracy forsubtitles is measured and considered as acceptable when subtitles achieve a score of at least 98% with the WER or NER model: this score is considered as the minimum accuracy requisite. Before examining the data of this study, an Inter-Annotator Agreement (IAA) test was carried out to validate the methodological framework and the taxonomic scheme described above. In computational linguistics and, in particular, in speech corpora analysis, the usage of annotation represents an important tool to analyse audio/video material and make specific comments or add detailed information on a set of texts (Bendazzoli 2010: 76). Yet, before continuing with the categorisation and analysis of 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)[9]

The taxonomy defined above should therefore be evaluated so as to assess whether it is reproducible by other annotators or evaluators – and hence sufficiently reliable. 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' (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 specifically, 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 (seven 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 (IAA) test. IAA is described by Gagliardi as follows:

[9] All translations are my own unless otherwise noted.

> Within the computational context, I.A.A. 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)

The external annotators involved in the testing phase included eight researchers/PhD students working and studying in the linguistic field, all coming from the Department of Interpreting and Translation (University of Bologna). The participants not directly engaged in the present research and included six female individuals and two male individuals with an age ranging from twenty five to fifty years old (with seven annotators of Italian nationality and one of Chinese nationality).

3. Results and discussion

After having presented the methodology at the basis of this study, the analysis of data is now possible. When examining the IAA test results, it is first of all possible to claim that a larger portion of the annotators involved identified the presence or absence of an error with respect to the given Perfect Matches (PM): 89% for the first file and 92.5% for the second file. Secondly, a high IAA rate was obtained for the taxonomic scheme Layer 1 (Coarse-Grained Error categories) of this study, for both files: 89% and 100% (including and excluding Null errors, respectively) with the first file and 92% and 98.30% with the second file. For the Fine-Grained Error taxonomy, the rates were lower and this was mostly due to a major ambiguity between pairs of categories and to the higher probability of entering a different value (as there are five different categories to choose from). However, these values of agreement remain substantial and can prove the validity of this taxonomic level too. Finally, when considering the error seriousness categorisation (the pair Serious/Not Serious), the IAA rate achieved a good level of agreement among the eight annotators, with values of 85% (first file) and 84.85% (second file). These results allowed considering this study’s taxonomic scheme to be as sufficiently reliable and reproducible, given the substantial levels achieved (to use the conceptual categorisation discussed in Fuoli and Hommerberg 2015: 334; Gagliardi 2018: 5).

Table 5 below shows the WER and NER rates calculated for all files. In particular, the Table includes the *Min.* and *Max.* values for all three rates, including the relevant *MEAN* values and the *Standard Deviation*. The data refer to both Native-speaker and Non Native-speaker files.

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 5. WER, NER1 and NER2 rates for all database files.

By subdividing the files into two groups (Non-Native and Native speakers) as shown in Tables 6 and 7 below, it is possible to observe that Native-speaker files reported 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 6. 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 7. WER, NER1 and NER2 rates for Non Native-speaker files.

More specifically, it is possible to specify that the mean values for Non-Native files (see Table 7 above) were of 92.31% (WER), 94.02% (NER1) and 95.79% (NER2), and they are all below the minimum accuracy requisite (i.e. 98%). On the other hand, with Native speaker files (Table 6 above), the accuracy rate was slightly higher: 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. 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), the files with WER and NER1 accuracy rates around 90% may however be considered as acceptable for the respeaking process (which is not incorporated into this study's investigations), 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 partial hearing loss, 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.

When comparing Google Speech Recognition (GSR) engine's output with that generated by VoxSigma, the following data can be obtained. The comparison was carried out for a limited number of files only.

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

Table 8. 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 9. Comparison of NER1 mean values between GSR engine and VoxSigma.

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

Table 10. 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 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.

From the analysis of data, another significant aspect emerged: the importance of terminology-related errors. In fact, from the data 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. One of the most important novelties of this study is probably the analysis of the impact of terminological resources on the processing of a ASR+NMT system, as well as on its 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 are mainly referenced to as ‘OOV – Out of Vocabulary’ errors. A mentioning of this feature is also reported in other studies from Romero-Fresco and other scholars (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, 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. 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. In a context-specific scenario like the international conferences on climate change, built-in terminological resources did in fact prove to be not sufficient. For this reason, a new concept of Augmented Terminology was introduced in ASR+NMT analysis and accuracy evaluation in order to properly cope with this challenge. To enhance ASR+NMT performances, the system’s terminology should in fact 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 organising the institutional communications. After incorporating the concept of Augmented Terminology, the pipeline for an efficient ASR+NMT system would therefore appear like the one represented below.

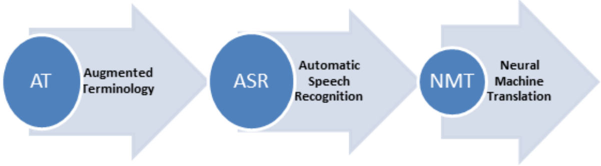

Figure 7. ASR+NMT system pipeline including Augmented Terminology.

To better understand Figure 7 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 organisation) 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 generation of subtitles in target languages.

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. 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 only, 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 terminological resources were downloaded from the Food and Agriculture Organisation's FAOTERM Portal[10]. The FAO office supplied a series of uploadable files (in particular, the IFADTERM, the Climate Change and Bioenergy database, the FAOTERM glossary and, finally, the Oceanography database) for the purposes of the experiment. All these databases were delivered in the *.xlsx* format (compatible with VoxSigma platform) and they were appropriately validated by the relevant organisation (i.e. the FAO). The analysis showed that most of the recognition errors encountered in previous processing were corrected. This operation permitted to obtain a higher accuracy in ASR for the file in question, taking the value of previous NER rate (95.60%) to 99.36% (AT-adapted NER rate), well above the minimum accuracy requisite set by the industry.

4. Conclusions

In relation to the results of the analysis, it is possible to maintain that the overall quality of the subtitles examined was evaluated as sufficiently accurate for the Native speaker files only. 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 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, accuracy was again well below the minimum accuracy requisite (albeit performing slightly better), except for the Native speaker files, where the software almost ap-

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

proached and overcome 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 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, 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 re-speakers in the production of live 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 this 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). Different would the case be of intralingual subtitling for people with a partial loss of hearing or with minor hearing difficulties. 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.

Additionally, as mentioned in the Introduction, considered the frequent, non-availability of interpreters (and re-speakers for non-hearing people) at the institutional level for any target language and language combination, the application of ASR technology, combined with NMT, may possibly help in breaking down the barriers of communication in the case of Native-speaker conferences within the global institutional context. Under these organisations, multilingualism is indeed a 'fundamental pillar of Institutional Translation' (Jopek Bosiacka, 2013) and ASR + NMT technology may contribute to preserve multilingualism.

Finally, a final consideration can be added in relation to this innovative approach involving terminology in the evaluation of accuracy. This study in fact showed that, with the application of Augmented Terminology resources, a higher accuracy can be obtained. 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. Additionally, this new concept finally brought to the proposal of defining an adapted version of the NER model based on a terminology categorisation of errors.

Bibliography

Accipio Consulting (2006). 'Tecnologie del linguaggio per l'Europa', *Report on the Technologies of language financed by the European Union*, http://www.tcstar.org/pubblicazioni/ITC_ita.pdf, ITC IRST. Trento, Tipolitografia TEMI.

Armstrong, S. (1997). 'Corpus-based methods for NLP and translation studies', *Interpreting* 2 (1–2), 141–162.

Bendazzoli, C. (2010). *Corpora e interpretazione simultanea*. Bologna: Asterisco.

Britannica, The Editors of Encyclopedia. (2020). 'Database'. *Encyclopedia Britannica*, <https://www.britannica.com/technology/database>, accessed 13 May 2021.

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)*. Agadir, Morocco, 2016, Piscataway, NJ: IEEE, pp. 223

European Union (2020). *Interinstitutional Style Guide*, <https://publications.europa.eu/code/en/en-000100.htm>, accessed 17 December 2020.

Fuoli, M. and Hommerberg, C. (2015). 'Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions', *Corpora*, 10, 315–349.

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

Goman, A. (2017). 'Addressing Hearing Loss With an Aging Population', *The Hearing Journal*, 70 (6), 6.

Greco, G. M. (2016). 'On Accessibility as a Human Right, with an Application to Media Accessibility'. In Matamala, A., and Orero, P. (eds.), *Researching Audio Description. New Approaches*, pp. 11–33. London: Palgrave.

Istituto per gli Studi di Politica Internazionale (ISPI) (2012). 'The long walk to gender parity in international organizations'. *Publications for the Italian Parliament and Ministry of Foreign Affairs*, <https://www.ispionline.it/it/pubblicazione/long-walk-gender-parity-international-organizations>, accessed on 10 May 2021.

Jopek Bosiacka, A. (2013). 'Comparative law and equivalence assessment of systembound terms in EU legal translation'. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 12, 110–146, <https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/237 >.

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 26–27*, pp. 58–65.

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, 11–12 April 2017*. Paris, Nanterre University, Slides: 29.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

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'. *Language & Communication*, 49, 56–69.

Romero-Fresco, P. (2018). 'In support of a wide notion of media accessibility: Access to content and access to creation', *Journal of Audiovisual Translation*, Vol. 1: pp. 187–204.

Romero-Fresco, P. and Martínez, J. (2015). 'Accuracy Rate in Live Subtitling: The NER Model'. In Díaz-Cintas, J., and Baños, R. (ed.), *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape*, pp. 28–50. London: Palgrave MacMillan.

Romero-Fresco, P. and Pöchhacker, F. (2017). 'Quality assessment in interlingual live subtitling: The NTR model'. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 16, 149–167.

Russo, M., Bendazzoli, C., Sandrelli, A., and Spinolo, N. (2012). 'The European Parliament Interpreting Corpus (EPIC): Implementation and developments'. In Straniero Sergio, F., and Falbo, C. (eds), *Breaking ground in corpus-based interpreting studies*, pp. 35–90. Bern: Peter Lang.

Saon, G. and Chien, J. (2012). 'Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances', *IEEE Signal Processing Magazine*, 29 (6), 18–33.

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, 15 March 2019*. Forlì, University of Bologna. Slides: 24.