D5.1 A TOOL ALLOWING POST-EDITING OF TEXT

(OTHER, PU, M15, STXT)

Revision: v.1.0

Work package	WP5
Task	Task 5.1
Due date	30/03/2022 → 30/06/2022
Submission date	13/06/2022
Deliverable lead	SWISS TXT
Version	1.0
Authors	Juan Martinez (SWISS TXT), Luca Marra (SWISS TXT)
Reviewers	Davy Van Landuyt (EUD) Mélanie Hénault-Tessier (INTerpretis)

Abstract	This document describes the features of the interface of the NERstar editor, a tool within EASIER that will allow human post-editors to review and evaluate Natural Machine Translation (NTM) of spoken language text in the context of sign language and audio descriptions, as well as of traditional, live, and automatic segmented text. The editor is provided as a web-based system that runs in Google Cloud, accessible via an URL.
Keywords	Editing, post-editing, text, segmented text, translation, editor, recognition

Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

DISCLAIMER

COPYRIGHT NOTICE

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:	**OTHER***	
Dissemination Level		
PU	Public, fully open, e.g. web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to EASIER project and Commission Services	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

Funded by the Horizon 2020 Framework Programme of the European Union

EXECUTIVE SUMMARY

This document describes the features of the interface of the NERstar editor, a tool within EASIER that allows human post-editors to review and evaluate Neural Machine Translation (NTM) of spoken language text in the context of sign language, as will detailed below, and audio descriptions, as well as of traditional, live, automatic texts, and segmented texts. Neural machine translation (NMT) is an approach to machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model. The editor is provided as a web-based system that runs in Google Cloud, accessible via an URL. This deliverable shows in detail:

1. A description of the NERstar model and the NER formula, which is at the core of the NERstar evaluation tool.
2. A description of the new editor and its editing functionalities.
3. An outline of the new approach to (segmented) text correction.
4. A description of the functionalities to edit and correct Automatic Speech Recognition (ASR) (segmented) text with NERstar.
5. Some concrete guidelines on correction and editing.
6. A summary of the findings for the implementation in SL.

By elaborating on these points, the document delivers a consistent framework and solid guidelines for generating impact in the standardization of automatic (segmented) text translations within EASIER that will be used for automatically recognized speech and the correction of its text output.

Funded by the Horizon 2020
Framework Programme of the European Union

NERSTAR EDITOR ACCESS INFORMATION

URL: https://easier.nerstar.online/

LOGIN: easier@nerstar.online

PSW: **Demo@0815**

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE OF CONTENTS

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES AND VIDEOS

Funded by the Horizon 2020
Framework Programme of the European Union

1 INTRODUCTION AND SCOPE OF THE DOCUMENT

This document describes the features of the interface of the NERstar editor, a tool within EASIER that allows human post-editors to review and evaluate Neural Machine Translation (NTM) of spoken language text in the context of sign language and audio descriptions, as well as of traditional, live, automatic texts, and segmented texts. The editor is provided as a web-based system that runs in Google Cloud, accessible via an URL. It opens new avenues for the correction of automatically recognized speech and its text representation.

Funded by the Horizon 2020
Framework Programme of the European Union

2 REMARKS

Over the past few years, Audiovisual Translation (AVT) seems to have shifted its focus from quantity to quality. In the case of live text production, and more specifically respeaking, the most common method to assess the quality of automatic texts and segmented texts produced in real-time is usually to assess their accuracy or lack thereof. In addition, other quality issues in this field concern the delay, position, character identification, speed, as well as features related to the perception of viewers, such as overall quality assessment, comprehension. The NERstar editor aims to solve and improve these quality issues and extend them to automatic produced sign language, text recognition, audio descriptions in EASIER.

Funded by the Horizon 2020
Framework Programme of the European Union

3 OBJECTIVES AND STRUCTURE OF THE DOCUMENT

The aim of this deliverable is to present an overview of the editor. The remainder of this document is organized as follows.

- In Chapter 4 we provide a description of the NERstar model and the NER formula, which is at the core of the NERstar evaluation tool.
- In Chapter 5 we describe the new editor and its editing functionalities.
- In Chapter 6 we outline the new approach to automatic texts and segmented texts correction.
- Chapter 7 describes the functionalities to edit and correct Automatic Speech Recognition (ASR) text segments with NERstar.
- Chapter 8 details concrete guidelines on correction and editing.
- Chapter 9 summarizes the findings.

Funded by the Horizon 2020
Framework Programme of the European Union

4 THE NER FORMULA

NER is an acronym for Number, Edition error, and Recognition error. The NER model, on which the NERstar editor is based, was initially created for subtitling. It is a method for determining the accuracy of live automatic texts and segmented texts, e.g. in television broadcasts and events, that are produced using speech recognition and is heavily viewer-centered. Live texts may be expected to reach 98% accuracy.

The model contains a formula to determine the quality of live texts: A NER value of 100 indicates that the content was coded 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 text. This number is then divided by the total number of words in the live texts and finally multiplied by one hundred.

The acronyms in the formula stand for the following:

1. N (number) = total number of words in the live texts
2. E (Edition error) = edition error
3. R (Recognition error) = recognition error

Below is an explanation of the three components of the formula:

N: Number of words in the respoken text, including commands (punctuation marks, speaker identification, etc.) and words.

E: Edition errors; usually caused by the strategies applied by the respeaker. In other words, they are the result of the respeaker's judgment or decision. The most common situation is when the respeaker omits or adds something because e.g., the original speech rate is too fast. This leads to the loss of a piece of information or the introduction of wrong information due to the miscomprehension of the original text.

Editing errors are calculated by comparing the respoken text and the original text and may be classified as serious, standard, or minor, scoring 1, 0.5, and 0.25, respectively. In the case of automatically produced text segments, edition errors are, among others, those related to incorrect capitalization, punctuation, and speaker identification.

R: Recognition errors; usually these are misrecognitions caused either by mispronunciations/mishearing or by the specific technology used to produce the texts. These errors may be insertions, deletions, or substitutions, and are calculated by comparing the respoken text and the original text. Again, they may be classified as serious, standard or minor, scoring 1, 0.5, and 0.25 respectively.

In addition, the model includes the number of Correct Editions (CE) and an assessment component. Correct editions are instances in which the respeaker's editing has not led to a loss of information, which is calculated by comparing the respoken text and the original text. Given the difficulty involved in producing verbatim live texts, the omission of redundancies and hesitations may be considered as cases of correct edition and not as errors if the coherence and cohesion of the original discourse are maintained.

The results are then analyzed in the assessment section. The assessment deals with several issues, including, but not limited to, the speed and delay of the automatically produced text segments, how the respeaker has coped with the original speech rate, the overall flow of the texts on the screen, speaker identification, the audiovisual coherence between the original

Funded by the Horizon 2020 Framework Programme of the European Union

image/sound and the texts, and possible loss of time in the corrections. This assessment determines the quality of texts in the NER model.

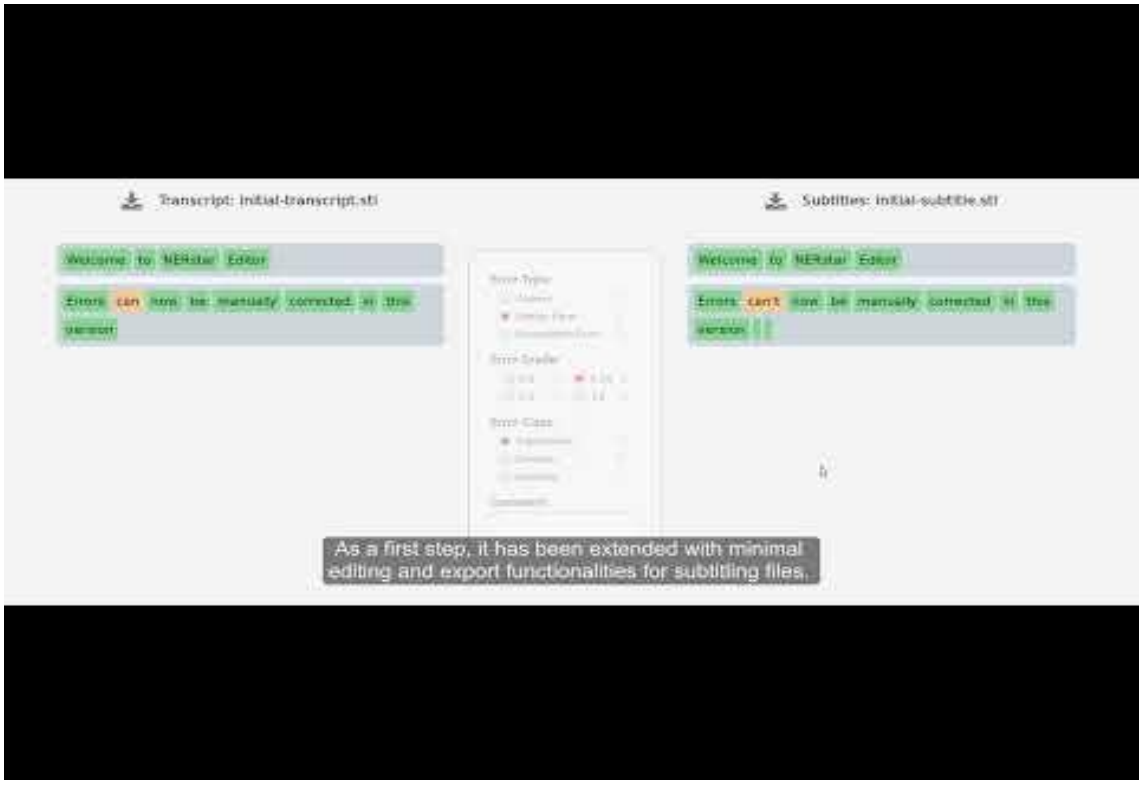
Funded by the Horizon 2020
Framework Programme of the European Union

5 THE NERSTAR EVALUATION TOOL

As a first step, the NERstar evaluation tool was developed to analyze and correct text created by an ASR in an easy and efficient way. Text segments are easy to catch by the eyes, to read and to use for further purposes. The implementation of NERstar within EASIER has in its first version minimal editing and export functionalities for automatically translated files. The next step will be to extend it in EASIER to process automatically recognized sign language and its text output.

As shown in the Video 1, the NERStar evaluation tool allows editors to post-edit automatically recognized text created by automatic speech recognition systems (ASR) as well as to start with the initial creation of text segments.

There are three strong arguments for the use of ASR systems for the correction or initial creation of text segments.



VIDEO 1: INTRODUCTION TO NERSTAR EDITOR
https://www.youtube.com/watch?v=qgwf2f4dqrg

From a productivity perspective, the use of ASR systems enables an increase in amount of text production.

From an application perspective, an ASR system does not “censor” as is the – well-intentioned – case with texts created manually or by respeaking. Manually created segmented texts (e.g., captions for TV) are normally summarized and reformulated. This has two important consequences: First, such an ASR system can reproduce 100% of what has been said. Second, the display of the individual words in the same order in which they are spoken enables lip-reading parallel to the reading of texts (*see Dhoest and Rijckaert 2021 on deaf people’s habits in news broadcasting [2]*).

Funded by the Horizon 2020
Framework Programme of the European Union

From a technical perspective, these last two points have wide implications for the improvement of the readability and comprehensibility of ASR texts. Using corrected text files has the potential to further improve the recognition performance of the ASR system, as well as the segmentation and time optimization algorithms developed by SWISS TXT within the EASIER project.

Funded by the Horizon 2020
Framework Programme of the European Union

6 NEW APPROACH TO SEGMENTED TEXT CORRECTION

First of all, the space character is equated with a single word:

- Like a single word, a space is given a time-in and a time-out code.
- A space can be – like a single word – edited, substituted or deleted.
- Corrections only affect text level, as well as time-in and time-out codes of single words; spaces are therefore not affected.

Following a NER evaluation of errors and tests run with other correction systems (*see Romero-Fresco and Martinez 2015 for details on the development of the editor [1]*), this new approach is justified by the fact that most of the time spent to correct texts is devoted to

- placing the cursor in certain places,
- typing short words that have to be corrected (e.g.: “the” instead of “their”) or rewritten because they are missing as they are not recognized by the ASR (e.g.: "to", "in”),
- deleting, controlling, or recreating superfluous spaces, and
- searching the video for individual words or word groups that were incorrectly recognized and re-annotating them.”

This has led to the development of the following functionalities in the NERstar editor:

- Correctors no longer use the cursor in the actual file, nor do they type in it. Single words or spaces can be marked for correction, which can then be edited in two correction windows or, ideally, substituted or deleted with one click.
- The same applies to the correction or creation of punctuation marks; these are no longer typed manually, but – ideally with one click – either edited/generated, substituted, or deleted by selecting spaces.
- A single click of a play button plays the part of the video that needs to be heard to ensure the recognized text is correct.

Funded by the Horizon 2020
Framework Programme of the European Union

7 FUNCTIONALITIES TO EDIT AND CORRECT AUTOMATIC SPEECH RECOGNITION (ASR) SUBTITLES

For the display of the work surfaces of the two editors (post-processing or first creation of texts), we used the surface of the NERstar tool as a basis – graphically as well as functionally.

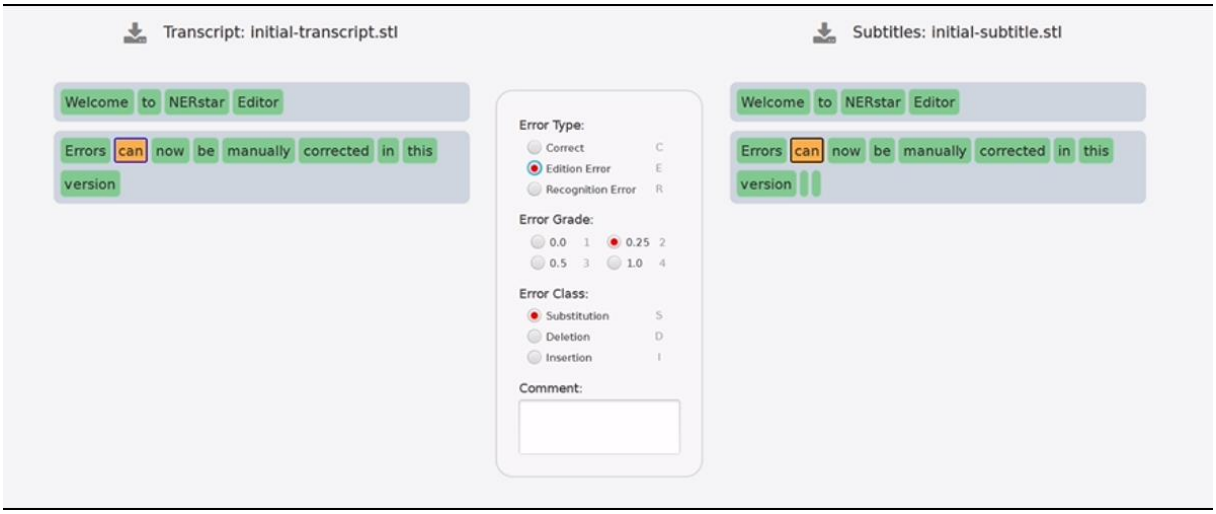


FIGURE 1: THIS FIGURE IS TAKEN FROM THE VIDEO "INTRODUCTION TO NERSTAR EDITOR"

Similar to the NERstar tool, we have equipped the two editors with editing, substitution, and deletion functionalities, both at the single-word and space-character levels. In both editors the dictionary and macros would be active in the background, making "reasonable" suggestions to be able to substitute incorrectly recognized individual words with one click.

Importantly, the current display of the text to be corrected does not correspond to scrolling text but displays the text exactly as it is shown in Video 1. The rationale behind this feature is clarity, as it is easier to read the text and spot mistakes when it is not scrolling. Also, a play button will be available in each text box for one-click listening to the desired text segment in the video or audio in case of doubts when correcting what was said.

This is shown in Video 2, "*Text editing and creation in NERstar for EASIER*", which exemplifies the correction flow for a German text. The interface comprises the video with text segments, the text to be corrected, as well as two windows for the correction of text or spaces/punctuation, respectively. Both the text and space windows show the current edition word/space, the possible substitutes offered by the dictionary (for words), and a deletion box. As shown, the editor allows for the correction of words, the addition/removal of spaces and punctuation marks to improve the clarity of the text, as well as the addition of functional words or the correct case marking to generate clean data for the further improvement of the ASR system.

Funded by the Horizon 2020 Framework Programme of the European Union

VIDEO 2: TEXT EDITING AND CREATION IN NERSTAR FOR EASIER
https://www.youtube.com/watch?v=XbGsLw7RNSU

During correction, the word order of automatically recognized single words is not modified to allow for

- enabling lip-reading and
- displaying everything that was said to deal with voluntary and involuntary omissions leading to meaning or information loss

The goal is primarily not to produce a grammatically correct text, but one that is “communicatively” correct with respect to

- what is said being left in the order in which it was spoken in the original comment to enable lip-reading and avoid”, and
- what is said only being corrected if a recognition error generated by the ASR changes the meaning of the original comment or represents a loss of important information.

7.1 FOCUS

The focus is on correcting ASR recognition errors (inserts, omissions, mix-ups), which are defined as "serious" or "normal" in the NER model. Basically, editors carry out a kind of NER evaluation of an automatically recognized text, in which not the grammatical correctness, but the adherence of the text to the original are the main criteria they use in their correction work.

7.2 PUNCTUATION

Punctuation marks structure a text, making it easier to understand. Incorrect or missing punctuation marks are classified as serious errors because they

Funded by the Horizon 2020
Framework Programme of the European Union

- could change the meaning of the original comment and create a new meaning that makes sense within the context in question, and
- could also be used as a means of controlling automatic segmentation.

7.3 TEXT CONSOLIDATION

So far, the handling of recognition errors defined as "minor" in the NER model, i.e., minimal omissions or insertions that are so minor that the correct word can be easily recovered, has not been addressed.

Although the effort required to correct a single minor error is reasonable, it is important to note that it is one of the most typical error categories when using ASR recognition systems. "Typical" here means that many such small errors in turn require a lot of time for correction.

In this context, we suggest checking the possibility of what can be called an “automatic text consolidation” before a manual correction. Namely, before a person starts the correction work, the recognized ASR text runs through a correction program so that minor recognition errors can be corrected automatically. These include, e.g., errors at pronoun or article level, case marking errors, capitalization, as well as missing or incorrect punctuation marks.

This will allow for editors to focus on correcting recognition errors on a single word or word group level that change the meaning of the original comment or represent a loss of important information, and not on those where it is about adding or deleting a single letter.

The workflow for the initial creation of text segments in an editor is similar to what is shown in video (2): The video to be transcribed is recognized by ASR: single word boxes, spaces, and punctuation and time codes are generated. Afterward, everything is corrected.

Funded by the Horizon 2020
Framework Programme of the European Union

8 CONCRETE GUIDELINES SAMPLE

Until an automatic text consolidation program is available to automatically correct minor recognition errors, as well as incorrect or missing punctuation marks, the following error categories in an ASR text file should be corrected manually, if possible, without changing the order of the recognized individual words:

1. misplaced or missing punctuation marks
2. «minor» and «normal» recognition errors breaking the communicative flow of information
3. individual words (proper names, key terms) and numbers important for the facts and understanding
4. single-word groups acting as units of meaning or – due to erroneous recognition – changing the meaning of the original comment
5. words representing the loss of important information

As mentioned in Chapter 7, the non-correction of grammatically incorrect structures or structures deviating from the standard variety of a language is justified by the fact that a correction would impair the possibility of lip-reading.

Funded by the Horizon 2020
Framework Programme of the European Union

9 KEY POINTS AND CONCLUDING REMARKS

This document provides an overview of the NERstar model and the new editor. We present a new approach to segmented text correction, and the new functionalities to edit and correct automatically generated subtitles based on ASR within EASIER. We use ASR systems because they do not allow for the loss of communicative content, allow for lip-reading and create data that, once corrected, can be used to improve the system. The interface also offers more options for corrections thanks to its editable nature. This has wide-ranging consequences for EASIER: it will be used for automatically produced sign language, text recognition, and audio descriptions. Future work will focus on improving automatic recognition for deaf signers and deaf translators. As is well known, a word-by-word translation is not ideal in the context of sign language, where the use of rephrasing abounds due to the linguistic structure of sign languages.

The goal of EASIER is to develop a fully automated, smart tool to enable deaf and hearing individuals to communicate and interact using sign language. To enable this goal, the post-editing environment enabled by Nerstar within EASIER to allow for the correction of machine-translated text and the improvement of automatic transcription we have shown in this document will be extended to sign language. In particular, we will support the post-editing of sign language animations to allow first for deaf translators to write glosses and in the future for glossing automatically recognized text. We will use spoken language to sign language translations rendered by an avatar: similar to what we have shown in this document with regard to text, the editor will display the animation on a timeline with the source text in parallel. When a signing error is found, the user can select the part to correct and record the corrected sign-language passage through the editor. As for text, the post-editing corrections will be collected to improve the accuracy of the tool.

Funded by the Horizon 2020
Framework Programme of the European Union

REFERENCES

[1] Romero-Fresco Pablo, Martinez Juan (2015) Accuracy Rate in Live Subtitling: The NER Model. In: Piñero, R.B., Cintas, J.D. (eds) Audiovisual Translation in a Global Context. Palgrave Studies in Translating and Interpreting. Palgrave Macmillan, London. https://doi.org/10.1057/9781137552891_3.

[2] Dhoest Alexander & Jorn Rijckaert (2021) News 'with' or 'in' sign language? Case study on the comprehensibility of sign language in news broadcasts, Perspectives, DOI: 10.1080/0907676X.2021.1936088.

Funded by the Horizon 2020
Framework Programme of the European Union