easier

intElligent Automatic
SIgn languagE tRanslation

D6.8 TOOLS FOR HARMONIZING BROADCAST TEXT DATA

Revision: v1.0

Work Package	WP6
Task	T6.3
Due date	30/06/2022
Submission date	30/06/2022
Deliverable lead	University of Surrey
Version	1.0
Authors	Richard Bowden, Ozge Mercanoglu Sincan (University of Surrey – UNIS)
Reviewers	Rosalee Wolfe, Eleni Efthimiou (ATHENA), Michael Filhol (CNRS)

Abstract	This deliverable gives an overview of the tools used to parse subtitles to obtain spoken language pairs for broadcast footage sign language.
Keywords	text resources, broadcast data, subtitles

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

Document Revision History

Version	Date	Description of change	List of contributors
V0.1	16/06/2022	First draft	Richard Bowden (UNIS)
V0.2	17/06/2022	Internal Review	Rosalee Wolfe, Eleni Efthimiou (ATHENA)
V0.3	22/06/2022	Internal Review	Michael Filhol (CNRS)
V0.4	22/06/2022	Second draft	Ozge Mercanoglu Sincan (UNIS)
V0.5	23/06/2022	Internal Review	Eleni Efthimiou (ATHENA)
V0.6	30/06/2022	Final Edit	Richard Bowden (UNIS)
V1.0	30/06/2022	Camera-ready submission	Giacomo Inches (Martel)

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the “Intelligent Automatic Sign Language Translation” (EASIER) project’s consortium under EC grant agreement 101016982 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable		**R**
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 deliverable reports work that is the output of Task 6.3, namely work relating to the processing of broadcast/subtitle data which took place during the period of Month 1-18 in the project EASIER. Task 6.3 of work package 6 is devoted to the processing of this broadcast data which includes video, audio, captions and subtitles. The focus of T6.3 was to provide a homogeneous data representation for use in the project. D6.7 provided an overview of the available formats and standards and tools were then created to convert each to a common format. D6.8 outlines these tools.

Funded by the Horizon 2020
Framework Programme of the European Union

CONTENTS

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

Sign Languages

BSL	British Sign Language
DGS	German Sign Language / Deutsche Gebärdensprache
DSGS	Swiss-German Sign Language / Deutschschweizer Gebärdensprache
LIS	Italian Sign Language / Lingua Italiana dei Segni
LSF	French Sign Language / Langue des Signes Française

Funded by the Horizon 2020
Framework Programme of the European Union

1 TOOLS FOR HARMONIZING BROADCAST TEXT DATA

This deliverable provides an overview of the text resources that were previously described in D6.7. The software tools that are used in the processing of extracting whole sentences from subtitles will be described. Finally, the VIA annotation tool to be used in the manual alignment of subtitles will be discussed.

1.1 SUBTITLE EXTRACTION

Deliverable D6.7 provided an overview of the available text resources captured and curated as part of D4.1. It also provided statistics which are summarized in Table 1.1. For each country, the broadcast footage contained picture-in-picture sign language interpretations, and the associated subtitles in the spoken language of that nation. This provides language pairs for British Sign Language (BSL) - English, German Sign Language (DGS) - German, Swiss-German Sign Language (DSGS) - Swiss-German, Italian Sign Language (LIS) - Italian and French Sign Language (LSF) - French.

Software tools for processing the data were based on open source solutions. We used the **FFMPEG** library to downloaded video and subtitle files. We kept the original formats of the subtitle files when downloaded, which are VTT for BSL-English, TTML for DGS-German and SRT for the rest of the language pairs.

Table 1.1: *Statistics of the downloaded subtitle files categorized with respect to the languages. Duration format is (hours:minutes:seconds).*

Sign Lang.	Spoken Lang.	# Videos	← Duration	# Subtitles	← Duration	Format
BSL	English	1,962	1467:22:56	1,962	1131:04:52	WebVTT
DGS	German	6,189	2162:25:09	4,913	1418:27:02	TTML
DSGS	Swiss-German	4,058	2454:20:35	2,126	834:17:39	SRT
LIS	Italian	1,260	949:33:23	249	99:21:44	SRT
LSF	French	1,770	955:09:53	907	420:01:16	SRT

We parsed the subtitle files in Python. For WebVTT format, we utilized the webvtt-py library (**webvtt-py**). For the SRT format, we used the pysrt library (**pysrt**). Finally, for the TTLM subtitles, we first convert them into SRT format using an open-source script (**ttml2srt**), and then utilized the pysrt library.

Detailed statistics for the data processed were provided in D6.7, but this yielded 3,257,796 sentences containing 34,543,446 individual word tokens over 5 spoken languages.

1.2 EXTRACTING SENTENCES FROM SUBTITLES

To extract spoken language sentences from the subtitle files, we developed a tool in Python, utilising the Natural Language Toolkit (**NLTK**). The aim was to extract the sentences with their beginning and end timings and create new subtitle files that include whole sentences for use in WP4.

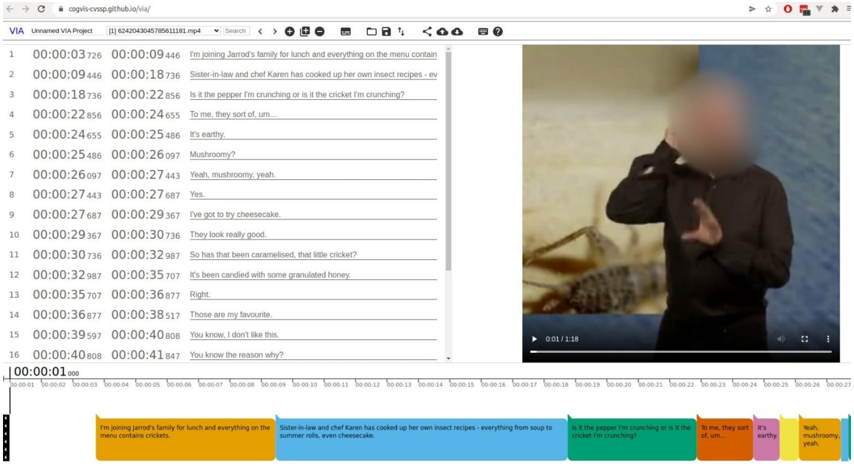
The tool expects a main directory path that contains subtitle files, and an output path that specifies where to save the results. Some of the DGS subtitles contain *<font></font>* tags. Therefore, the script firstly eliminates these tags. Then, it uses the *sent_tokenize* and *word_tokenize* functions, which provide a tokenizer at the level of sentences and words. When the script detects the end of a sentence within a subtitle, it calculates the duration of the sentence and estimates the end/start time according to the number of words. Finally, new subtitle files are created from the assembled sentences and their timings.

We executed our tool on all aforementioned datasets, i.e., BSL, DGS, DSGS, LIS, and LSF. The original subtitles of the BSL dataset contain whole sentences. Therefore, we compared the number of sentences of the processed BSL subtitles with the original ones as a sanity check. Our tool produced the same number of sentences in the majority of cases. We found the failure cases to be caused by the use of multiple punctuation in a sentence.

1.3 ALIGNING SENTENCES TO VIDEO

The VGG Image Annotator (**VIA**) is an open source, manual annotation software tool for image, audio and video data. Within the scope of the EASIER project, we will use VIA tool to manually align audio-aligned subtitles to video. A version can be seen running at `https://cogvis-cvssp.github.io/via/`.

In order to annotate a video, both the video and subtitle file are loaded into the tool. The interface shows the subtitles on a timeline and allows the user to add, delete, edit or adjust the timing of each subtitle. Fig. 1.1 shows the interface of the VIA tool. Once the annotation is completed, the updated subtitles can be exported as either a VTT or CSV file.



1	00:00:03 726	00:00:09 446	I'm joining Jarrod's family for lunch and everything on the menu contain
2	00:00:09 446	00:00:18 736	Sister-in-law and chef Karen has cooked up her own insect recipes - ev
3	00:00:18 736	00:00:22 856	Is it the pepper I'm crunching or is it the cricket I'm crunching?
4	00:00:22 856	00:00:24 655	To me, they sort of, um...
5	00:00:24 655	00:00:25 486	It's earthy.
6	00:00:25 486	00:00:26 097	Mushroomy?
7	00:00:26 097	00:00:27 443	Yeah, mushroomy, yeah.
8	00:00:27 443	00:00:27 687	Yes.
9	00:00:27 687	00:00:29 367	I've got to try cheesecake.
10	00:00:29 367	00:00:30 736	They look really good.
11	00:00:30 736	00:00:32 987	So has that been caramelised, that little cricket?
12	00:00:32 987	00:00:35 707	It's been candied with some granulated honey.
13	00:00:35 707	00:00:36 877	Right.
14	00:00:36 877	00:00:38 517	Those are my favourite.
15	00:00:39 597	00:00:40 808	You know, I don't like this.
16	00:00:40 808	00:00:41 847	You know the reason why?

I'm joining Jarrod's family for lunch and everything on the menu contains crickets.

Sister-in-law and chef Karen has cooked up her own insect recipes - everything from soup to summer rolls, even cheesecake.

Is it the pepper I'm crunching or is it the cricket I'm crunching?

To me, they sort of, um...

It's earthy.

Yeah, mushroomy, yeah.

Figure 1.1: *The interface of the VIA annotation tool.*

Library	**Source**
FFMPEG	https://ffmpeg.org/
webvtt-py	https://pypi.org/project/webvtt-py/0.1/
pysrt	https://github.com/byroot/pysrt
ttml2srt	https://www.npmjs.com/package/ttml2srt
NLTK	https://www.nltk.org/natural-language-toolkit
VIA	https://www.robots.ox.ac.uk/ vgg/software/via

Funded by the Horizon 2020
Framework Programme of the European Union

2 CONCLUSION

In this deliverable, we first gave an overview of the available text resources captured as part of D4.1. We developed a Python tool to create new subtitles which include sentences as a whole with their beginning and end timings. The newly created subtitles will be used in WP4. We then gave brief information about a VIA annotation tool, which is suitable for aligning subtitles manually.

Funded by the Horizon 2020
Framework Programme of the European Union