easier
intElligent Automatic
SIgn languagE tRanslation

D6.7 OVERVIEW OF TYPES OF TEXT RESOURCES IN BROADCAST DATA

Revision: v1.0

Work Package	WP6
Task	T6.3
Due date	31/01/2022
Submission date	28/01/2022
Deliverable lead	University of Surrey
Version	1.0
Authors	Necati Cihan Camgoz, Richard Bowden (University of Surrey – UNIS)
Reviewers	Onno Crasborn (Radboud University – RU), Kearsy Cormier (UCL Deafness, Cognition and Language Research Centre – DCAL)

Abstract	This deliverable gives an overview of the subtitles curated by the Deliverable 4.1. Available subtitles are summarized and categorised with respect to the corresponding spoken language. These subtitles are then parsed and processed to obtain sentence, word and vocabulary statistics of available text resources.

Keywords	text resources, broadcast data, subtitles, word statistics

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	18/01/2022	First draft	Necati Cihan Camgoz (UNIS)
V0.2	19/01/2022	Editing	Richard Bowden (UNIS)
V0.3	23/01/2022	Applying Onno Crasborn's Comments	Necati Cihan Camgoz (UNIS)
V1.0	23/01/2022	Final Corrections	Necati Cihan Camgoz (UNIS)

DISCLAIMER

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 the work done in Task 6.3, which involves processing and harmonization of text resources in broadcast data. The datasets curated in Deliverable 4.1 were downloaded and the available text resources inspected and categorised with corresponding spoken language. In this deliverable we give details of the available text resources and share the language statistics extracted from the parsed subtitles.

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 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 AVAILABLE TEXT RESOURCES IN BROADCAST DATA

In this section we provide an overview of the available text resources contained in the broadcast datasets curated by D4.1, , statistics of which are summarized in Table 1.1. The crawled datasets contain URLs for video files, which contain picture-in-picture sign language interpretations, and subtitles, which have timed text information for the spoken language aspects of the broadcast. The available sign language - spoken language pairs are: 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.

Using the FFMPEG library, we downloaded video and subtitle files. We kept the original formats of the subtitle files while downloaded, which are VTT (W3C, 2019) for BSL-English, TTML (W3C, 2020) for DGS-German and SRT (Matroska, 2020) for the rest of the language pairs.

The statistics of the downloaded subtitle files and their formats, along with the available videos are shared in Table 1.1. As can be seen, and highlighted in D4.1, there is a discrepancy between videos and subtitles. Due to the crawled URL-based nature of D4.1 and the online data expiration policies of the broadcasters, some of the subtitle URLs were not valid at the time of download. BSL-English sequences did not suffer from this issue, as they were obtained directly from the BBC as a public dataset (Albanie et al., 2021).

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

Funded by the Horizon 2020 Framework Programme of the European Union

2 LANGUAGE STATISTICS OF THE PARSED SUBTITLES

To have a better understanding of the size and the breadth (domain-wise) of the available text resources, in this section we share language statistics obtained from the parsed subtitles of D4.1.

Using the tools developed in T6.3, we parsed the subtitles in python. For WebVTT format, we utilized webvtt-py library (glut23, 2020). For SRT format, we used the pysrt library (byroot, 2022). Finally, for TTLM subtitles, we first convert them into SRT format using an open-source script (codingcatgirl, 2018), and then utilized the pysrt library.

To estimate the number of samples (aligned spoken language-sign video pair) which we can generate from D4.1, we first extract the number of available spoken language sentences. We apply NLTK library's language specific tokenizers (Bird et al., 2009) to segment subtitles into sentences for each file. This yield us 3,257,796 sentences containing 34,543,446 individual word tokens over 5 spoken languages.

We also wanted to investigate the breadth of the domains the text resources cover. Since the manual annotation of dataset is not feasible, we used word statistics as a proxy. We split sentences into individual words and calculated the number of unique tokens, which is 575,953 over the 5 languages. One interesting finding was that German and Swiss-German languages have a higher unique token ratio compared to other languages. We believe this is due to how German and Swiss-German compounds are treated in the orthography, how they are written: as a single string without a space between the compound parts. Other languages, such as English, are also rich in compounds, but separate the compound parts by a space. Decomposing German and Swiss-German compounds might be necessary while matching them to co-articulated signs. We also extracted the number of rare words and calculate the number of 'singletons', words that only occur once, and 'rares', words that occur less than five times. Over five languages 30-47% of all the vocabulary items only occur once in the dataset, and 55-73% of the unique tokens occur less than five times. The detailed statistics categorized in terms of spoken language can be seen in Table 2.1. English has a lower rare word percentage compared to other languages. The difference between English, Italian and French vs German and Swiss-German suggest this may be due to compounding, However, this may also be due to the different types of broadcast footage used while curating these datasets. While other languages mainly contain news footage, which can have a higher number of unique words, such as entities, dates and places, the English language samples are episodes from a wide variety of TV series, which may share common vocabulary (Inches and Crestani, 2013).

Table 2.1: *Word statistics of the text resources from different languages.*

Language	# Tokens	# Sentences	Vocab. Size	# Singletons (=1)	# Rares (<5)
English	11,797,594	1,147,263	80,146	24,328 (0.30)	43,971 (0.55)
German	11,360,616	1,133,266	222,635	102,136 (0.46)	160,317 (0.72)
Swiss-German	7,351,780	646,153	175,992	83,430 (0.47)	128,221 (0.73)
Italian	811,329	56,419	36,785	15,528 (0.42)	25,745 (0.70)
French	3,221,531	274,008	61,820	21,881 (0.35)	38,185 (0.62)

Funded by the Horizon 2020 Framework Programme of the European Union

REFERENCES

Albanie, Samuel, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. (2021). "BBC-Oxford British Sign Language Dataset". In: *arXiv preprint arXiv:2111.03635*.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

byroot (2022). *pysrt*. URL: https://github.com/byroot/pysrt (visited on 01/18/2022).

codingcatgirl (2018). *ttml2srt*. URL: https://github.com/codingcatgirl/ttml2srt (visited on 01/18/2022).

glut23 (2020). *webvtt-py*. URL: https://github.com/glut23/webvtt-py (visited on 01/18/2022).

Inches, Giacomo and Fabio Crestani (2013). "An introduction to the novel challenges in information retrieval for social media". In: *PROMISE Winter School*. Springer, pp. 1–30.

Matroska (2020). *SRT Subtitles*. URL: www.matroska.org/technical/subtitles.html#srt-subtitles (visited on 01/18/2022).

W3C, World Wide Web Consortium - (2019). *WebVTT: The Web Video Text Tracks Format*. URL: https://www.w3.org/TR/webvtt1/ (visited on 01/18/2022).

— (2020). *Timed Text Markup Language 1 (TTML1) (Third Edition)*. URL: https://www.w3.org/TR/2018/REC-ttml1-20181108/ (visited on 01/18/2022).

Funded by the Horizon 2020 Framework Programme of the European Union