


Camomile project

Corpus selection

Contractual Deliverable D4.1

(Target) Release version: V1.0

Dissemination Level		
PU	Public (can be made available outside of Camomile Consortium without restrictions)	
RE	Restricted to Camomile participants and a specified group outside of Camomile consortium	
CI	Camomile Internal (only available to (all) Camomile participants)	X
CL	Camomile Limited (only available to a specified subset of Camomile participants)	
Distribution list (only for RE or CL documents)		

0 General Information

0.1 Document

Title	Corpus Selection
Type	Contractual Deliverable
Ref	D4.1
Target version	V1.0
Current issue	V0.1
Status	Draft
File	`livrable.4.1.tex`
Author(s)	Gilles ADDA / CNRS-LIMSI
Reviewer(s)	Claude Barras / CNRS-LIMSI
Release date	<Dd/mm/yyyy>

0.2 History

Date	Version	Comment
09/08/2013	V0_1	

0.3 Document scope and structure

This document describes the different corpora that will be used during the Camomile project.
Contributors:

CNRS-LIMSI: Gilles Adda, Claude Barras

ITU: Hazim Kemal Ekenel

UPC: Ramon Morros, Javier Hernando

0.4 Content

Show	Train	Dev	Test
BFM Story	7:57:49	1:00:50	0:59:48
Culture et Vous	2:09:28	0:15:00	0:15:03
Ça vous regarde	2:00:05	0:15:39	0:15:01
Entre les lignes	1:59:43	0:15:00	0:15:02
Pile et Face	2:01:26	0:15:04	0:15:01
LCP Info	4:07:09	0:30:08	0:29:56
Top Questions	3:57:41	0:30:02	0:27:01
Total	24:13:23	3:01:46	2:56:55

Table 1: TV shows currently present in the corpus

1 Executive Summary

We have decided to use four different corpora, two in French language, one in Catalan, and one in Turkish. The main corpus is the one used in the REPERE evaluation[1]; we plan to enhance the face annotation of this corpus. The other French corpus is the Canal 9 Political Debates[2]. The two non-French video corpora are included mainly to assess the portability of the tools to another languages, and to experiment with collaborative annotation when some sites/annotators do not speak the language of the video. The annotations of the two non-French corpora will be done according to the use-cases in which they will be involved; therefore the exact definition of these annotations will be decided when the different use cases will be decided.

In the following sections, we will first present the two corpora in French language, then the two corpora in Turkish and in Catalan

2 REPERE Corpus: people recognition in multimodal conditions

2.1 The REPERE evaluation

Finding people on video is a major issue when various informations come from television and from the Internet. The challenge is to understand how to use the information about people that comes from the speech and the image and combine them so as to determine who is speaking and who is present in the video.

Started in 2011, the REPERE Challenge [2] aims to support the development of automatic systems for people recognition in a multimodal context. Funded by the French research agency (ANR) and the French defense procurement agency (DGA), this project has started in March 2011 and ends in March 2014. To assess the systems' progress, the first of two international campaigns has been organized at the beginning of 2013 by the Evaluation and Language resources Distribution Agency (ELDA) and the Laboratoire national de métrologie et d'essais (LNE). The second official campaign is open to external consortia who want to participate in this challenge and will take place at the beginning of 2014.

2.2 Sources

The January 2013 corpus represented 24 hours of training data, 3 hours of development data and 3 hours of evaluation data and is described in Table 1. The videos are selected from two French TV channels, BFM TV and LCP, for which ELDA has obtained distribution agreements. The shows are varied: *Top Questions* is extracts from parliamentary "Questions to the government" sessions, featuring essentially prepared speech.

Ça vous regarde, Pile et Face and *Entre les lignes* are variants of the debate setup with a mix of prepared and spontaneous but relatively policed speech.

[1] http://www.defi-repere.fr/
[2] http://www.idiap.ch/scientific-research/resources/canal-9-political-debates/

Figure 1: Some example frames from the video corpus

LCP Info and *BFM Story* are modern format information shows, with a small number of studio presenters, lots of on-scene presenters, interviews with complex and dynamic picture composition.

Culture et vous, previously named *Planète Showbiz*, is a celebrity news show with a voice over, lots of unnamed known people shown and essentially spontaneous speech.

These video were selected to showcase a variety of situation in both the audio and video domains. A first criteria has been to reach a fair share between prepared and spontaneous speech. A second one was to ensure a variety of filming conditions (luminosity, head size, camera angles...). For instance, the sizes of the heads the annotators would spontaneously segment varied from 146 pixels^2 to 96,720 pixels^2 for an image resolution of 720x576. Some example frames are given Figure 1.

2.3 Annotations

Two kinds of annotations are produced in the REPERE corpus : audio annotation with rich speech transcription and video annotation with head and embedded text annotation.

2.3.1 Speech annotations

Speech annotation are produced in `trs` format using the Transcriber software [1]. The annotation guidelines are the ones created in the ESTER2 [3] project for rich speech transcription. The following elements are annotated :

- Speaker turn segmentation.
- Speaker naming.
- Rich speech transcription tasks gather segmentation, transcription and discourse annotation (hesitations, disfluences...)
- The annotation of named-entities of type “person” in the speech transcription with a normalized label for each identity.

Figure 2: Segmentation example

2.3.2 Visual annotations

In complement to the audio annotation, the video annotation has necessitated the creation of specific annotation guidelines. The VIPER-GT video annotation tool has been selected for its ability to segment objects with complex shapes and to enable specific annotation schemes. The video annotations consist in the six following tasks:

- Head segmentation: all the heads that have an area larger than 1000 pixels2 are isolated. Heads are delimited by polygons that best fit the outlines. Figure 2 is an example of head segmentation. It is worth noting that it is head segmentation and not face segmentation. Sideways poses are annotated too.
- Head description: each segmented head may have physical attributes (glasses, headdress, mustache, beard, piercing or other). The head orientation is also indicated: face, sideways, back. The orientation choice is based on the visible eyes count. Finally, the fact that some objects hide a part of the segmented head is indicated, specifying the object's type.
- People identification: The name of the people is indicated. Only well-known people and the people named in the video are annotated. Unknown people have are identified with a unique numerical ID.
- Embedded text segmentation and transcription: the transcription of the segmented text is a direct transcript of what appears in the video. All characters are reproduced with preservation of capital letters, word wrap, line break, etc. Targeted texts are segmented with rectangles that fit best the outlines (see figure 2). Also whether a text is part of an identification cartouche is also annotated.
- Named-entities (type "person") annotation in transcripts of embedded texts
- The annotation of appearance and disappearance timestamps: the aim is to identify the segments where the annotated object (head or text) is present.

2.3.3 Global annotations

Beyond the parallel annotation of audio and visual content, the corpus creation pays special attention to the multimodal annotation consistency. A people names database ensures the coherence of given names in audio and visual annotations. Moreover, unknown people IDs are harmonized when the same person appears both in audio and video annotations.

		Train	Dev	Test
Visual	Heads seen	13188	1534	2081
	Words seen	120384	14811	15844
Speech	Segments	12833	1602	1514
	Words	275276	34662	36489
Persons	Seen known	725	146	141
	Speaking known	556	122	126
	To find	811	172	162
	Seen unknown	1907	238	160
	Speaking unknown	1108	163	179
	Names on screen	729	138	160
	Names cited	870	190	161
Clues modalities	Name appears	504	83	83
	Name cited	544	116	101
	Never named	178	39	36
	Not speaking	255	50	36
	Not seen	86	26	21
	Speaking and seen	470	96	105

Table 2: the REPERE first evaluation corpus

In addition two per-person annotation are provided for both video and audio: the gender of the person, and its role in the show under a 5-class taxonomy.

2.4 First evaluation corpus

Table 2 summaries the annotations done on the 30 hours of corpus created for that run, and the number of persons that can be found through audio or visual clues. We can see that in the test corpus 51% of the people to find have their name appearing on screen and 62% are introduced in the speech. In practice the OCR is much more reliable than the speech recognition for proper names, making these 51% is primary information source for the global system. Interestingly, 22% of the persons are never named, limiting the reachable level for unsupervised systems. A number of persons appear only in one modality. In the test 22% are only visible, which is a little lower than in the rest of the corpus, and 13% are only heard.

3 Canal9: Swiss French speaking political debates

3.1 Presentation

The corpus distributed by IDIAP institute includes 70 recordings for a total of 43 hours and 10 minutes of material [4]. Each debate revolves around a yes/no question like “Are you favorable to new laws on scientific research?”. The participants state their answer (yes or no) at the beginning of the debate and do not change it during the discussion. Each debate involves a moderator that tries to give the same space to all participants (or at least to the two fronts corresponding to yes and no supporters). Furthermore, the moderator tends to reduce tensions when the discussion becomes too heated.

3.2 Format and Structure

The recordings are available as high-quality full-frame (720 Œ 576 pixels) DV compressed PAL recordings, along with an uncompressed audio stream sampled at 48 kHz. They have been live edited and not all the participants are visible all the time. All debates took place in the same recording studio (with no audience) and Figure 3 shows some of the most frequent camera views: full group (19.7% of data time), personal shots (66.1% of data

Figure 3: Most frequent camera views from Canal9 corpus.

time), and multiple participants (11.0% of data time). The remaining 3.2% corresponds to short reports (typically at the beginning of the debate) and credits shown at both beginning and end of each debate.

3.3 Group Composition

Political debates include two main roles: moderator and participant.

3.3.1 The Participants

The spatial arrangement of the participants reflects the situation (see full group view in Figure 3). The two factions physically oppose one another in a spatial arrangement that has been shown to elicit agreement between people on the same side and disagreement between people on opposite sides. Overall there are 190 unique participants, 154 participate only in one debate, 25 participate in two debates, and the remaining 11 participate in three. In terms of gender, the set of the participants includes 25 women and 165 men.

3.3.2 The Moderator

All debates include one moderator expected to ensure that all participants have at disposition the same amount of time for expressing their opinion. Furthermore, the moderator intervenes whenever the debate becomes too heated and people tend to interrupt one another or to talk together. Overall, there are five different moderators, 1 woman and 4 men. The woman moderates 28 debates, while the men moderate 24, 9, 8 and 1 debates, respectively.

3.4 Duration Distribution

In total, the 70 debates of the corpus correspond to 43 hours, 10 minutes and 48 seconds. Of these, 41 hours 50 minutes and 40 seconds (96.9% of the total) correspond to actual discussions, while the remaining time includes reports and credits shown at beginning and end of each debate.

3.5 Annotations

3.5.1 Manual Speaker Segmentation

The audio of each debate has been manually segmented into single speaker intervals. Speakers are identified with a label that does not correspond to their names, and all the turns (single speaker segments) where the same person talks hold the same label. The segmentations are stored as `trs` files, an XML format used by the publicly available transcriber annotation tool.

3.5.2 Role

The annotations report the role played by each person involved in the debates, i.e. moderator (the journalist expected to guarantee that all persons have enough time to express their opinion and that tries to inhibit aggressive and impolite behaviors) or participant (the persons that support one of the two answers to the question around which the debate revolves).

3.5.3 Agreement and Disagreement

The participants are labeled in terms of *group-1* and *group- 2* according to how they answer to the central question of the debate. Participants belonging to the same group agree with one another, while participants belonging to different groups disagree with one another.

3.5.4 Automatic Speaker Segmentation

The output of an automatic speaker diarization system is available for the audio channel of each debate. The segmentations are available as `trs` files.

3.5.5 Manual Shot Segmentation

The video channel of each debate is manually segmented into shots, i.e. time intervals between two changes of camera. The shot segmentation is available as a list of shot boundaries, i.e. time instants where the camera changes. The boundaries are stored in ASCII files.

3.5.6 Automatic Shot Segmentation

The output of an automatic shot segmentation system is available for the video channel of each debate. The format of the automatic shot segmentations is the same as the one of the manual ones.

3.5.7 Manual Shot Classification

Each shot is annotated in terms of two classes: personal shot (see Figure 3) and other. No automatic classification is available.

3.5.8 Manual Identification of Participants in Personal Shot

All personal shots showing a given participant are annotated with her/his identity. No automatic version of this annotation is available.

3.6 Availability

The corpus is publicly available through the web-portal of the Social Signal Processing Network [3] upon signature of an appropriate End User Licence Agreement.

[3] `www.sspnet.eu`

4 Turkish video corpus

The corpus consists in the recording of a 1h30 of a TV channel in Turkish. The corpus has been recorded by ITU.

The contents of this corpus is described below :

10.18 - 10.26 newscast for afternoon

10-26 - 10.30 ads

10.30 - 10.48 discussion program (only two people that present the program are speaking. No sound for background images)

10.48 - 10.51 ads

10.51 - 10.54 weather conditions (no images of people)

10.53 - 10.55 ads

10.55 - 11.24 newscast (images of famous people from politics (Turkish) and sports)

11.24 - 11.28 ads (known people and their sounds are present in movie trailers for Turkey)

11.28 - 11.49 discussion program about real estate (only identities and voices of 2 people)

11.49 - 11.50 ads (voices and pictures of people from politics but just for 1-3 seconds.)

No annotation is done.

5 TV3 Agora: Catalan video corpus

The corpus consists in news programs (TV3 AGORA), in which two or more people are discussing about a subject around a table (see Figure 4 for some images of the set). The corpus has been recorded by UPC.

The program starts with a black screen. Then the host spawn on closeup and describes the program, sometimes with someone else at the background. Then, they show a general view of all people present on the set (between 5-7 people), and when the discussion starts they show a closeup on the person speaking. The break matches the commercial break, and after that the discussion continues shooting three different scenes: a general view, a closeup and a medium shot with 3-4 people, on most of them will be side-viewed or back. A total of 126 different people are present in the videos. Around 10 people are present in more than one video, at different times.

There are 67 videos, each one is about 40 mn long; thus the total size is about 44 hours of video. The total size is 53.1 GB in VOB format and 10.9 GB in mp4.

No annotation is done.

[1] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber:ădevelopment and use of a tool for assisting speech corpora production. *Speech Communication*, 33:5–22, 2001. Special issue on speech annotation and corpus tools.

[2] Olivier Galibert and Juliette Kahn. The first official REPERE evaluation. In *Proceedings of The First Workshop on Speech, Language and Audio in Multimedia (SLAM)*, 22-23 August 2013.

Figure 4: Some images extracted from the TV3 Agora corpus

[3] Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *INTERSPEECH*, pages 1149–1152. ISCA, 2005.

[4] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin. Canal9: A database of political debates for analysis of social interactions. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (IEEE International Workshop on Social Signal Processing)*, 2009. Publication Date: 10-12 Sept. 2009.