Title	Overview of multimodal corpora in the CLARIN infrastructure
Version	0.5
Author(s)	Darja Fišer, Jakob Lenardič
Date	09-09-2020
Status	Draft
Distribution	BoD, NCF, UI
ID	CE-2020-1737

Table of contents

1. Introduction

In the following report, we present an overview of multimodal corpora that are part of the CLARIN infrastructure (i.e., they are either listed in the VLO or in the repositories of the national consortia). Following Abuczki and Baiat Ghazaleh (2013), we understand multimodal corpora in a fairly narrow sense, as data collections used to study how two or more modalities interface with one another in human communication. In this sense, multimodal corpora are often collections of video and speech recordings accompanied with transcriptions and gesture annotations,[1] although multimodal corpora of textual data supplemented with images exist as well (Allwood 2008: 209). Such corpora can be used for "the exploration of a range of lexical, prosodic and gestural features of conversation, and for investigations of the ways in which these features interact in real, everyday speech" (Abuczki and Baiat Ghazaleh 2013: 88).

The report was conducted in three steps:

(i) by manually searching the VLO and the national consortia with the following keywords: "video corpus", "multimodal corpus", "image corpus", and "gesture corpus", as well as with the simple keywords "image", "gesture", and "video" and narrowing the results down to corpora using faceted search
(ii) with input provided by CLARIN UI and NC coordinators.

The full results are available in this Google Docs Spreadsheet. In total, 16 corpora were identified in the survey. In Section 2, we first provide a comprehensive list of the multimodal corpora that are part of the CLARIN infrastructure. In Section 3, describe their identification (i.e., listed in the VLO or not), their availability (download or through a concordancer), and their metadata (language, size, research discipline and period of publication, annotation, and license). Section 4 concludes the report with a discussion of the metadata of the corpora both from the quantitative and qualitative points of view.

2. The multimodal corpora in the CLARIN infrastructure

16 multimodal corpora have been identified in the CLARIN infrastructure. In what follows, we group them according to language (monolingual and multilingual corpora) and describe them with the following metadata:

(i) Size
(ii) Annotation
(iii) Licence
(iv) Language
(v) Availability (online querying and/or download)
(vi) Publication

[1]Conversely, general speech corpora that only include recordings and transcriptions aren't "multimodal corpora" in this narrow sense, as the transcriptions and recordings semiotically relate to one and the same modality i.e. that of spoken language. For an overview of speech corpora in the CLARIN infrastructure, see the Spoken corpora resource family.

In what follows, we present 14 multimodal corpora that include video and audio recordings (Section 2.1.) and 2 corpora that present the multimodal interaction between texts and accompanying images (Section 2.2.).

2.1. Video-audio corpora

Table 1: Video-audio multimodal corpora in the CLARIN infrastructure, sorted by language

Corpus	Language	Description
IFA Dialog Video corpus **Size:** 5 hours **Annotation:** functional annotation of dialogue utterances, annotated gaze direction **Licence:** GNU general public license	Dutch	This corpus contains annotated video recordings of friendly Face-to-Face dialogues. It is modelled on the Face-to-Face dialogues in the Spoken Dutch Corpus (CGN). The procedures and design of the corpus were adapted to make this corpus useful for other researchers of Dutch speech. For this corpus 20 dialogue conversations of 15 minutes were recorded and annotated, in total 5 hours of speech. To stay close to the Face-to-Face dialogues in the CGN, pairs of well-acquainted participants were selected, either good friends, relatives, or long-time colleagues. The participants were allowed to talk about any topic they wanted. The corpus is available for download from a dedicated webpage (hosted by CLARIAH-NL). For a relevant publication, see van Son et al. (2008).
MPI ESF Corpus	Dutch, English, French, German, Swedish	This corpus was built under the ESF Foreign Language Speakers project. It contains a lot of annotated audio recordings containing multimodal interaction.
Eye-tracking in Multimodal Interaction Corpus **Licence:** restricted	English	The corpus is available for download from the Language Archive (CLARIAH-NL).
TV News Corpus **Size:** 30 hours **Licence:** CC-BY-SA	Estonian	This corpus contains video and audio recordings and their transcriptions. The corpus is available for download from META-SHARE (CELR distribution).
Corpus d'interactions dialogales **Size:** 8 hours	French	A demo version of this corpus is available for download (videos and transcriptions) from the ORTOLANG repository.

Annotation: prosody, interpausal units, gestures, syntax		For a relevant publication, see Bertrand et al. (2008).
BAS SmartWeb Video **Size:** 36 hours **Annotation:** orthography, phonology, speaker turn, noise, prosody, gaze direction **Licence:** CLARIN ACA	German	The corpus contains a collection of user queries to a naturally spoken Web interface with the main focus on the soccer world series in 2006. The recordings include 156 field recordings using a hand-held UMTS device (one person, SmartWeb Handheld Corpus SHC), 99 field recordings with video capture of the primary speaker and a secondary speaker (SmartWeb Video Corpus SVC) as well as 36 mobile recordings performed on a BMW motorbike (one speaker, SmartWeb Motorbike Corpus SMC). The corpus is available for download from the BAS CLARIN-D repository. For a relevant publication, see Mögele et al. (2006).
Natural Media Motion-Capture Corpus **Size:** 3 hours **Annotation:** gesture types, meta-information about encoding (e.g., difficult to encode) **Licence:** CLARIN ACA	German	This corpus contains data from 18 participants, whose task was to describe nine objects each to an experimenter, without using everyday vocabulary about forms, sizes or objects. The participants were recorded on audio and several video cameras, and their hand movements were recorded using an optical VICON motion capture system. The corpus is available for download from the BAS CLARIN-D repository.
SmartWeb Video Corpus (SVC) **Annotation:** speech segmented, video signal is spatially segmented for face detection **Licence:** ELRA (restricted)	German	This corpus contains 99 recordings of a human-human-machine dialogue: one speaker (which is being recorded) interacts with a human partner as well with a dialogue system via a smart phone (SmartWeb system). The speaker uses a client-server based dialogue system (SmartWeb) for spoken access to Internet contents in a natural environment (office, hallway, street, park, cafe, etc.). The corpus is available for download from the ELRA repository.
Bielefeld Speech and Gesture Alignment Corpus	German, English	This corpus contains 25 dialogues of interlocutors (50), who engage in a spatial communication task combining direction-

Size: 9881 isolated words, 1764 gestures **Annotation:** alignment of speech and gestures **Licence:** CLARIN ACA		giving and sight description. The stimulus is a model of a town presented in a Virtual Reality (VR) environment. Upon finishing a "bus ride" through the VR town along five landmarks, a router explained the route as well as the wayside landmarks to an unknown and naive follower. The corpus is available for download from the BAS CLARIN-D repository. For a related publication, see Lücking et al. (2013).
Multimodal and multiparty corpus of text comprehension interactions **Annotation:** orthographic transcription, gaze/head/eye/lip movements **Licence:** CC BY-NC-SA	Greek	This corpus contains reading comprehension exercises in a high school setting involving 2 high school students and their teacher. The goal of the sessions is to represent how the interaction between a teacher and more than one students is performed: what is the structure of the conversation; how turn-taking is coordinated; what are the multimodal feedback and attention signals the speakers employ. The corpus is available for download from CLARIN:EL. For a relevant publication, see Koutsombogera et al. (2016).
Hungarian Multimodal Corpus **Size:** 50 hours **Annotation:** non-verbal and verbal elements of communication **Licence:** open and restricted	Hungarian	This corpus contains video and audio recordings of conversations divided into two major parts: a simulated job interview and a guided dialogue about personal topics. The participants are university students (54 females, 67 males) mostly involving the same interviewer in both scenarios. The corpus is available for online browsing through the MTA RIL Language Archive Serve (HUN-CLARIN distribution) and for download from the Language Archive (CLARIAH-NL). For a relevant publication, see Pápay et al. (2011).
Multimodal corpus EVA 1.0 **Size:** 57 minutes	Slovenian	This corpus contains one episode of an audio/video session plus corresponding orthographic transcriptions with a duration of

Annotation: MSD-tagged, non-verbal and verbal elements of communication **Licence:** CC BY-NC-SA 4.0		57 minutes. The multi-party spontaneous discourse in the recording is from an entertaining evening TV-talk show A si ti tut not padu, broadcasted by the POP-TV Slovene commercial TV station in 2008, and represents a part of the Slovene spoken corpus GOS. In addition to the original transcription and morphosyntactic annotation from the GOS corpus, the following layers of information are added: • statement sentiment • phrase breaks within statements • prominence of statements • sentences within the statement • sentence sentiment • sentence type • speaker visibility on the scene • gesture units • gesture phrases • emotions • semiotic intent • dialogue role The corpus is available for download from the CLARIN.SI repository. For a relevant publication, see Mlakar et al. (2019).
Video-linked Thai/Swedish child data corpus **Annotation:** video-transcription alignment, word segmentation, phonetic transcription	Swedish, Thai	This corpus consists of 60 transcripts from interactions in everyday contexts between 6 children and their caregivers (10 transcripts per child), recorded longitudinally, for the period when the children are 18 to 27 months of age. All six children are growing up in middle class environments, in Sweden and Thailand (Bangkok area) respectively. The videos of the corpus are linked to the transcripts, on an utterance-by-utterance basis using the software CLAN (MacWhinney 2020). The corpus is available for online browsing (CLARIN K-Centre Lund University Humanities Lab). For a relevant publication, see Zlatev et al. (2006).
Unisa isiZulu Video Corpus	Zulu	The corpus is unavailable.

2.2. Text-image corpora

Table 2: text-image multimodal corpora in the CLARIN infrastructure, sorted by language

Corpus	Language	Description
A Multimodal Corpus of Tourist Brochures Produced by the City of Helsinki, Finland (1967-2008) **Size:** 58 double pages[2] **Annotation:** content, layout, graphic, typographic appearance, rhetorical structure **Licence:** CLARIN ACA	Finnish	This corpus contains tourist brochures produced by the city of Helsinki, Finland, is fully annotated using XML schema provided for the Genre and Multimodality (GeM) model (Bateman 2008). The corpus is available for download from FIN-CLARIN.
Hindi Visual Genome 1.0 **Size:** 32,925 items, 32,535 images, 32925 sentences, 322,000 words **Licence:** CC BY-NC-SA 4.0	Hindi, English	This corpus contains short English segments (captions) from Visual Genome along with associated images. The English texts are automatically translated to Hindi with manual post-editing, taking the associated images into account. The corpus is available for download from the LINDAT repository. For a relevant publication, see Parida et al. (2019).

[2]This refers to the fact that the tourist brochures in the brochures are scanned in sets of double pages.

3. Overview of the multimodal corpora

3.1. Identification

Only 1 (6%) of the 16 corpora cannot be found in the VLO:

(1) Multimodal and multiparty corpus of text comprehension interactions (CLARIN:EL)

3.2. Availability

3.2.1. For online querying and download

Only 1 (6%) corpus is available both for online querying and download:

(1) Hungarian Multimodal Corpus

The corpus is available for download under both open and restricted access from the Language Archive hosted by CLARIAH-NL. It can also be queried online through the Language Archive hosted by HUN-CLARIN.

3.2.2. For online querying

Only 1 (6%) corpus is available for online querying:

(1) Video-linked Thai/Swedish child data corpus

The corpus can be queried online through the corpus server of the Lund University Humanities Lab, which is a CLARIN K-Centre.

3.2.3. For download

The following 12 (75%) corpora are available only for download. In the parentheses, we specify the repository from which the corpus can be downloaded.

(1) Multimodal corpus EVA 1.0 (CLARIN.SI)
(2) Eye-tracking in Multimodal Interaction Corpus (CLARIAH-NL Language Archive)
(3) Corpus d'interactions dialogales (ORTOLANG)
(4) SmartWeb Video Corpus (SVC) (ELRA)
(5) Bielefeld Speech and Gesture Alignment Corpus (CLARIN-D BAS repository)
(6) Natural Media Motion-Capture Corpus (CLARIN-D BAS repository)
(7) TV News Corpus (CELR)
(8) Hindi Visual Genome 1.0 (LINDAT)
(9) IFA Dialog Video corpus (LINDAT; accessible through a dedicated webpage hosted by CLARIAH-NL)
(10) BAS SmartWeb Video (CLARIN-D BAS repository)
(11) A Multimodal Corpus of Tourist Brochures Produced by the City of Helsinki, Finland (1967-2008) (FIN-CLARIN)
(12) Multimodal and multiparty corpus of text comprehension interactions (CLARIN:EL)

Note that in the case of corpus (3), Corpus d'interactions dialogales, only a small demo subset can be downloaded from ORTOLANG. This fact is not stated anywhere in the repository.

3.2.4. Unavailable

The following 2 (13%) corpora are not available.

(1) Unisa isiZulu Video Corpus (SADiLaR)
(2) MPI ESF Corpus (LINDAT)

It is unclear why Unisa isiZulu Video Corpus is not available, although it should be noted that the corpus is poorly described in the repository. In the case of MPI ESF Corpus, the Project URL (https://archive.mpi.nl/?openpath=MPI556280%23) does not lead to the corpus but rather to the base landing page of the MPL Archive, compounding the accessibility.

3.3. Metadata

3.3.1. Language

The majority (12 or 75% out of 16) of the corpora are monolingual. Among the monolingual corpora, there are 3 German corpora, and 1 corpus per the following language: English, Estonian, Finnish, French, German, Greek, Hungarian, Slovenian, and Zulu.

The 4 multilingual corpora are the following. In the parentheses, we specify the languages in the corpus:

(1) Bielefeld Speech and Gesture Alignment Corpus (English, German)
(2) Hindi Visual Genome 1.0 (English, Hindi)
(3) Video-linked Thai/Swedish child data corpus (Swedish, Thai)
(4) MPI ESF Corpus (Dutch, English)

3.3.2. Size

Information on size is available for 10 (63%) out 16 corpora. It is worth noting that, out of all the resource families, multimodal corpora fare the worst with respect to including this information. The following 6 corpora lack this information:

(1) Eye-tracking in Multimodal Interaction Corpus
(2) SmartWeb Video Corpus (SVC)
(3) Unisa isiZulu Video Corpus
(4) Video-linked Thai/Swedish child data corpus
(5) MPI ESF Corpus
(6) Multimodal and multiparty corpus of text comprehension interactions

From a qualitative perspective, the corpora that report size do so unevenly. Most of the video-audio corpora that report size do so exclusively in terms of the length of the video-audio recordings, but omit information on the size of the included transcriptions in terms of words/tokens or other observed phenomena, such as the number of gestures. These are the following 7 corpora:

(1) Hungarian Multimodal Corpus (50 hours)
(2) Multimodal corpus EVA 1.0 (57 hours)
(3) Corpus d'interactions dialogales (8 hours)
(4) Natural Media Motion-Capture Corpus (3 hours)
(5) TV News Corpus (30 hours)
(6) IFA Dialog Video corpus (5 hours)
(7) BAS SmartWeb Video (36 hours)

By contrast, one corpus – namely, the Bielefeld Speech and Gesture Alignment Corpus – reports size in a more fine-grained manner, listing the length of the recordings (1 hour), the size of transcriptions (9881 isolated words), and the number of gestures (1764) annotated in the corpus.

Finally, one of the text-image corpora (cf. Section 2.2.) provides fine-grained information on size, including – aside from word number – the number of images, while the other corpus only reports the number of the included "double pages", but provides no information on e.g. token size or image number.

3.3.3. Annotation

Information on annotation is available for 11 (69%) of the 16 corpora. The following 5 corpora lack this information:

(1) Eye-tracking in Multimodal Interaction Corpus
(2) TV News Corpus
(3) Hindi Visual Genome 1.0
(4) Unisa isiZulu Video Corpus
(5) MPI ESF Corpus

The 11 corpora that describe annotation are quite variedly annotated, so we provide the full list here and specify the annotation layers in the parentheses. The most recurring annotation layers are gestures (4 corpora) and gaze direction (3 corpora).

(1) Hungarian Multimodal Corpus (*non-verbal and verbal elements of communication*)
(2) Multimodal corpus EVA 1.0 (*morphosyntactic annotation, gestures, statement/sentiment sentiment, speaker visibility*)
(3) Corpus d'interactions dialogales (*prosody, interpausal units, gestures, syntax*)
(4) SmartWeb Video Corpus (SVC) (*speech/video segmentation for face detection*)
(5) Bielefeld Speech and Gesture Alignment Corpus (*alignment of speech and gestures*)
(6) Natural Media Motion-Capture Corpus (*gesture types, meta-information about encoding (e.g., difficult to encode)*)
(7) IFA Dialog Video corpus (*functional annotation of dialogue utterances; annotated gaze direction*)
(8) Video-linked Thai/Swedish child data corpus (*video-transcription alignment, word segmentation, phonetic transcription*)
(9) BAS SmartWeb Video (*orthographic and phonologic transcription, annotation of speaker turn, noise, prosody, gaze direction*)
(10) A Multimodal Corpus of Tourist Brochures Produced by the City of Helsinki, Finland (1967-2008) (*content, layout, graphic, typographic appearance, and rhetorical structure*)
(11) Multimodal and multiparty corpus of text comprehension interactions (*orthographic transcription, gaze/head/eye/lip movements*)

Qualitatively, not all corpora describe the annotation layers at the same level of detail, with the descriptions ranging from vague (e.g., "non-verbal and verbal elements of communication") to detailed (cf. Multimodal corpus EVA 1.0).

3.3.4. Licence

12 out of 16 (75%) corpora contain information on licence. The following corpora lack this information:

(1) Corpus d'interactions dialogales
(2) Unisa isiZulu Video Corpus
(3) MPI ESF Corpus
(4) Video-linked Thai/Swedish child data corpus

The most common licences are as follows:

- CC-BY (5 corpora)
- CLARIN ACA (4 corpora)

4. Conclusion

In this report, we gave an overview of 16 multimodal corpora in the CLARIN infrastructure. We have presented their identification (i.e., whether they have VLO entries) and their availability (for download, online querying or both), as well as 5 types of metadata – language, size, publication period, annotation, and licence.

In terms of identification, 1 (6%) out of the 16 identified corpora is not listed in the VLO. However, identifying multimodal corpora in the narrow sense, i.e., corpora that are not just speech corpora including audio files and transcriptions, is hard in the VLO. For instance, the "multimodal" value listed under the Modality facet is underutilised, listing 6 entries, only 1 of which is a corpus.

In terms of availability, 2 (13%) out of the 16 corpora are unavailable for download and online querying; in both cases, unavailability is unclear and is compounded by out-of-date repository entries. Otherwise, availability is as follows: 1 (6%) corpus is available both for online querying and download, 12 (75%) corpora are available only for download, and 1 (6%) corpus is available only for online querying only.

In terms of language, most (12 or 75%) corpora are monolingual. All corpora save for 1 corpus of the Zulu language cover languages spoken in Europe, with German being the most represented language. The 4 multilingual corpora also contain non-European languages such as Thai and Hindi.

Information on size is available for all the 10 (63%) out of the 16 corpora. It is worth noting that, out of all the resource families, multimodal corpora fare the worst with respect to including this information. Furthermore, the corpora report size unevenly, with many corpora reporting only the length of the video/audio recordings but not the size of the included transcriptions. There are also some non-standard size types reported, such as the number of pages but no token/word count.

Information on annotation fares slightly better and is available for 11 (69%) of the 16 corpora. The annotation layers are varied, but not all corpora describe them at the same level of detail, some opting for very under-informative descriptors. Licence is available for 12 (75%) corpora, with CC-BY (5 corpora) and CLARIN ACA (4) being the most frequent types.

Lastly, the information provided in the description fields of many of the corpora in the repositories is generally poorer in comparison with the other resource families. Many corpora feature descriptions of the funding projects rather than the content of the corpus (e.g., Unisa isiZulu Video Corpus), while others feature sparse and often arbitrary bullet-point-style listings of the corpus's content (e.g., MPI ESF Corpus). Consequently, it is often difficult to tell whether a corpus is truly multimodal in the narrow sense (see Section 1) or whether multimodality only refers to the fact that the corpus is multimodal in the inclusion of more than one data type, as in the case of speech corpora which in addition to textual data also contain audio recordings, but aren't really "multimodal" like the corpora described in this report.

5. References

Abuczki, Ágnes, and Esfandiari Baiat Ghazaleh. 2013. An overview of multimodal corpora, annotation tools, and schemes. *Argumentum*, 9: 86–98. http://argumentum.unideb.hu/2013-anyagok/kulonszam/01_abuczkia_esfandiari_baiat.pdf.

Allwood, Jens. 2008. Multimodal corpora. In *Corpus linguistics: an international handbook (Vol 1)*, edited by A. Lüdeling and M. Kytö, 207–225. https://hal-hprints.archives-ouvertes.fr/hprints-00511882/document.

Bateman, John A. 2008. *Multimodality and Genre*. London: Palgrave Macmillan.

Bertrand, Roxane, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy. 2008. Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Autoatique des Langues*, 49 (3): 105–134. https://hal.archives-ouvertes.fr/hal-00349893.

Koutsombogera, Maria, Miltos Deligiannis, Maria Giagkou, and Harris Papageorgiou. 2016. Towards Modelling Multimodal and Multiparty Interaction in Educational Settings. In *Toward Robotic Socially Believable Behaving Systems*, edited by A. Esposito and L. Jain, vol. 106. https://doi.org/10.1007/978-3-319-31053-4_10.

Lücking, Andy, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2012. Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *J Multimodal User Interfaces*, 7: 5–18. https://doi.org/10.1007/s12193-012-0106-8.

MacWhinney, Brian. 2020. Tools for Analyzing Talk Part 2: The CLAN Program. https://talkbank.org/manuals/CLAN.pdf.

Mlakar, Izidor, Darinka Verdonik, Simona Majhenič, and Matej Rojc. 2019. Towards Pragmatic Understanding of Conversational Intent: A Multimodal Annotation Approach to Multiparty Informal Interaction – The EVA Corpus. In *SLSP 2019: Lecture Notes in Computer Science*, edited by C. Martín-Vide, M. Purver, and S. Pollak, 19–30. https://doi.org/10.1007/978-3-030-31372-2_2.

Mögele, Hannes, Moritz Kaiser, and Florian Schiel. 2006. SmartWeb UMTS Speech Data Collection. The SmartWeb Handheld Corpus. In *Proceedings of LREC2006*, 2106–2111. https://www.aclweb.org/anthology/L06-1151.

Pápay, Kinga, Szilvia Szeghalmy, and István Szekrényes. 2011. HuComTech Multimodal Corpus Annotation. http://argumentum.unideb.hu/2011-anyagok/works/PapayK_SzeghalmySz_SzekrenyesI.pdf.

Parida, Shantipriya, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. https://arxiv.org/abs/1907.08948.

van Son, R. J. J. H., Wieneke Wesseling, Eric Sanders, and Henk van den Heuvel. 2008. Promoting *free* Dialog Video Corpora: The IFADV Corpus Example. In *International LREC Workshop on Multimodal Corpora: MMCorp 2008: Multimodal Corpora*, edited by M. Kipp et al., 18–37. https://doi.org/10.1007/978-3-642-04793-0_2.

Zlatev, Jordan, Mats Andrén, and Soraya Osathanonda. 2006. A video-linked Thai/Swedish child data corpus: A tool for the study of comparative semiotic development. https://www.researchgate.net/publication/237333171_A_video-linked_ThaiSwedish_child_data_corpus_A_tool_for_the_study_of_comparative_semiotic_development.