Your Data at the Center of Attention: A Metadata Session Profile for Multimodal Corpora

Farina Freigang[1], **Matthias A. Priesters**[1], **Rie Nishio**[2], **Kirsten Bergmann**[1]
[1]Faculty of Technology, Center of Excellence "Cognitive Interaction Technology" (CITEC)
Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany
[2]Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Binderstr. 34, 20146 Hamburg, Germany
[1]{firstname.lastname}@uni-bielefeld.de; [2]rie.nishio@sign-lang.uni-hamburg.de

Keywords: Metadata profile, multimodal data, multimodal corpora, gesture, sign language, CMDI, ISOcat, CLARIN

1. Introduction

The production of high-quality multimodal corpora is extremely expensive and hence it is of major importance to manage these resources in a way that they are easily searchable and reusable for other researchers. In fact, the reuse of resources is an issue strongly promoted by research funding organizations, e.g., by the European Union in terms of their "open data strategy".[1] In the field of corpus linguistics and language resources it is widely agreed that the ever-expanding number and growth of corpora needs *metadata* for the purpose of corpus management. For linguistic resources there already exists a large number of metadata descriptions and metadata schemes, but so far not much effort has been put into the development of metadata descriptions and schemes for the particular structure of multimodal corpora.

This is, at least in parts, due to the fact that multimodal corpora are highly heterogeneous. They might include different modalities such as gestures, facial expressions, body posture or eye gaze for which no standardized coding schemes exist. Moreover, multimodal corpora might comprise multiple synchronous data streams, such as video, audio, time series data (e.g., motion capture or eye tracking) and annotation data. These aspects are not captured by recent metadata profiles. In CLARIN-D, the discipline-specific working group on "Speech and Other Modalities" has recently initiated a discussion on these issues (cf. Freigang and Bergmann, 2013) which has led to the proposal of a novel metadata session profile for multimodal data: the `MultimodalSessionProfile`.[2] The profile is based on a detailed evaluation of three different multimodal corpora.[3] It has been developed according to the CMDI standard (Broeder et al., 2012; de Vriend et al., 2013) including unique ISOcat[4] definitions within and for (but by no means exclusively for) the CLARIN[5] ERIC[6] infrastructure. It offers a wide variety of corpus descriptions especially designed for, but not limited to, multimodal data. This paper aims to present the new metadata session profile for multimodal data. In section 2., we start with a review of existing metadata schemes for multimodal data. In section 3. we introduce the `MultimodalSessionProfile` and in section 4. the corresponding `media-corpus-profile` is described, which has been developed in cooperation with the Bavarian Archive for Speech Signals[7] (BAS) as a multimodal extension of an earlier version thereof. We conclude with a discussion about metadata profiles in section 5.

2. Related work

In Freigang and Bergmann (2013), we compared relevant CMDI metadata profiles from the CLARIN Component Registry: `media-corpus-profile`, `media-session-profile`, `MultimodalCorpus`, and `BamdesMultimodalCorpus`. Those profiles lack, for instance, multimodal data descriptions such as for gesture annotations. Some modality components exist (`cmdi-modality`, `ModalityInfo`, etc.), however, their granularity is not fine enough (for a detailed discussion, see Freigang and Bergmann (2013)).

We identified two major problems with these profiles. First, when generating metadata descriptions for the previously mentioned multimodal corpora from existing metadata profiles, we found that the *granularity* in which modality metadata descriptions were possible, firstly, for certain annotation types and, secondly, for the study design and scenario of multimodal data was not fine enough. It was not possible to specify, for example, that iconic gestures are annotated in the data or the handedness of an actor[8]. Hence, from the

[1]`http://ec.europa.eu/digital-agenda/en/open-data-0`

[2]Monospaced font for designations denotes names of CMDI profiles/components/elements as they appear in the CLARIN Component Registry.

[3]Speech and Gesture Alignment ("SaGA") Corpus from Bielefeld University (Lücking et al., 2013), Dicta-Sign DGS Corpus from University of Hamburg (Matthes et al., 2012) and Natural Media Motion Capture ("NM-MoCap") Corpus from RWTH Aachen University (Hassemer, 2014).

[4]`http://www.isocat.org`

[5]CLARIN: Common Language Resources and Technology Infrastructure: `https://www.clarin.eu`

[6]ERIC: European Research Infrastructure Consortium

[7]`http://www.en.phonetik.uni-muenchen.de/research/bav_arch_spsig`

[8]We chose to use the term *actor* throughout our profiles and

corpus user's perspective, it was not possible to search for detailed features of multimodal corpora.
A second problem with existing metadata profiles was that technical descriptions were missing. With the recording of multimodal data, novel technical devices typical for gesture or sign language studies are used. One of our reference corpora includes, e.g., motion capture recordings of gestures. The `media-session-profile`, compared to other profiles, is rather advanced and already includes components for time series data and stereo video (3D) recordings. However, describing a marker setup as used in the NM-MoCap corpus in this metadata structure proved cumbersome and unintuitive. Therefore, a more meaningful way for describing motion capture data among others was one of the requirements for a new profile. Furthermore, descriptions for used technologies, as for example HD videos, were not elaborate enough and needed extension.

3. Introducing the `MultimodalSessionProfile`

Based on the identified problems, we developed various new components covering different modality aspects and also technical descriptions. As also discussed in Freigang and Bergmann (2013), there are several options of how to realize new metadata components. Since a lot of changes have been made, the integration of the new components into an existing profile structure was not feasible. Therefore, we created our own **`MultimodalSessionProfile`**[9] in a bottom-up fashion: After the design of new components, we added relevant existing components from the CLARIN Component Registry and integrated all into a large profile structure. The profile construction is oriented at `media-session-profile` by BAS und NaLiDa's[10] `MultimodalCorpus` profile (among others). The new components cover different modality aspects and also technical descriptions and most components are optional, meaning that the users can choose the relevant components themselves. In cooperation with BAS, we also created a multimodal version of the `media-corpus-profile`, discussed in section 4.

3.1. Modality components

We refined the granularity of metadata descriptions in two ways. First, in order to describe the modalities of the *annotations* in more detail and, second, to depict the details of used modalities by the *actor* and the modalities connected to the *study* design and scenario.

this paper, firstly, because it is most accurate. *Participant* and *subject* are terms which imply an arranged setting such as in studies, which is not always the case since some corpora are collections of data, such as the Dicta-Sign DGS Corpus or a collection of news broadcasts. Secondly, *actor* is the most neutral term available: the term *speaker* would exclude sign language and *signer* would exclude spoken language. Therefore, we used the term *actor* in newly created components of our session (and corpus) profile. The terms *subject* and *participant* occur rarely and only where components were reused.

[9] `http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1381926654659`

[10] `http://www.sfs.uni-tuebingen.de/nalida`

Multimodal annotations In the natural sciences and humanities, categorizing data is crucial for getting an overview and making sense of the data, and there are various, discipline specific ways of realising it. As for multimodal communication data, the categorization of observed phenomena is usually done by annotating recorded data accordingly: we capture category information in the description of the annotation schemata.
In gesture studies, for example, gestures can be classified according to different criteria. One popular method follows McNeill (1992), who defines *iconic* (resembling the content of speech), *metaphoric* (image of abstract concept), *deictic* (pointing) and *beat* (marking the structure of the utterance) gesture categories. Furthermore, McNeill temporally segments gestures into phases such as *preparation*, *stroke*, *hold* and *retraction*.

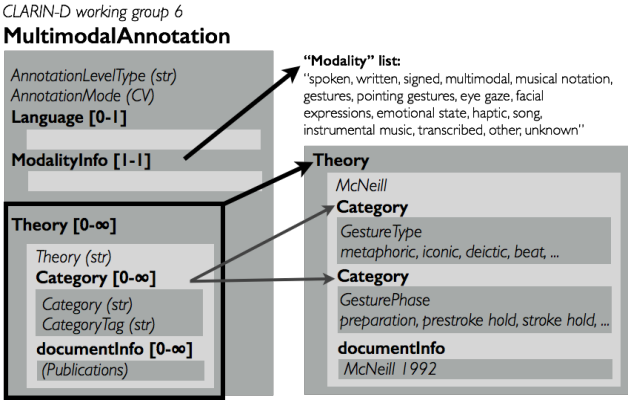

Figure 1: The `MultimodalAnnotation` component with an exemplary use of the `Theory` component for gesture categories by McNeill (1992).

In figure 1, we have sketched how a gesture annotation scheme based on McNeill's categories could be realized in our `MultimodalAnnotation` component. This component is kept simple in its design and is still flexible enough to cover complex category systems, also those which may be developed in the future. It can contain multiple `Theory` components which are given names and contain `Category` components. Categories are also named (e.g. "GestureType" and "GesturePhase") and contain `CategoryTag` elements, which represent the individual annotation labels (e.g. "iconic" or "preparation" in the above example). We explicitly encourage metadata creators to use this component also to refer to their own theory or annotation framework. Each `Theory` component can be enriched with literature references in the `documentInfo` component reused from the META-SHARE metadata profile (Gavrilidou et al., 2012). Additional information, e.g. explanations about the exact meaning of annotation categories, can be stored in optional `Description` elements. The theory component appears next to two other components and two elements in the `MultimodalAnnotation` component. One component is the modality list `ModalityInfo` mentioned in section 2., also developed by META-SHARE. It provides modality-related keywords for characterising the annotations performed on the

corpus data. The difference between the Theory and the ModalityInfo components is that the former describes the annotation scheme in detail, whereas the latter generally lists the modalities which were annotated. Furthermore, the metadata creator can specify the type of annotation level (e.g., *part of speech*, *gesture form*, etc.), the annotation mode (e.g., *manual*, *automatic*, etc.), and the language of the annotation.
We aimed at separating the annotation theory and the corresponding annotation labels from the technical information of the annotation file. The annotation file will be dicussed in subsection 3.2.

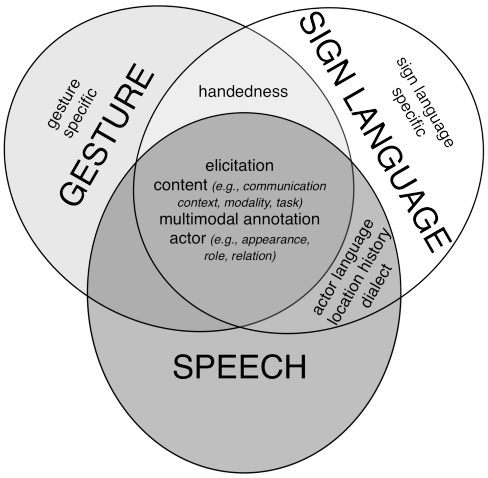

Figure 2: Newly created modality components classified into three main categories. Reused components are not indicated.

Actors Providing detailed information about the persons appearing in the corpus material was one goal of our metadata profile. Besides general information about actors (ActorPersonal), we also provide room for information which is more specifically needed for corpora involving gesture or sign language. For corpora including sign language, the personal history of actors, such as their educational background or the location where they grew up, are especially important, as these strongly influence sign language proficiency and the signed dialect. We built upon and extended a set of ISOcat data categories for describing signed language resources compiled and implemented by Crasborn and colleagues (Crasborn and Hanke, 2003a; Crasborn and Hanke, 2003b; Crasborn and Windhouwer, 2012). Furthermore our profile includes information about a actor's Handedness (either self-reported or assessed using a test), about the proficiency of spoken or signed languages (ActorLanguages) or about the relations between several persons (ActorRelation).

Design of data collection Modality aspects about the actors and the study as such were missing in the existing components. We defined new components that can be classified into three major categories: *gesture*, *sign language* and *speech* (figure 2). We developed category-specific descriptions, namely the ActorGestureSpecific and ActorSignLanguageSpecific components. Speech-specific descriptions are covered by reused components such as cmdi-subjectlanguages. Other components are kept general and allow for speech and sign language descriptions, among others, as does the component ActorLanguages. Other components that serve both of these categories are ActorDialect with its component LocationHistory. The component Handedness has been kept general for the description of both gesture and sign language. Finally, many components serve all three categories: The MultimodalElicitation component (which is supplemented by the reused Elicitation component for experimental research data) and, for example, the Content component comprising the study task, the modalities which are used during the study, and the communication context.

3.2. Technical metadata

Media files The profile includes fine-grained description categories for various types of media data, that is, *video*, *audio*, *image* and *time series* data. Most categories were reused from existing metadata profiles (most notably the media-session-profile), but some components were extended. Among the added features are information about camera perspectives, video dubbing/subtitling and the ability to describe multiple channels of one video recording (needed for 3D stereo video). The time series component was extended by components for marker sets used in optical motion capture systems and for kinematic data computed from raw motion capture data.

Annotation files The treatment of annotation files differs from existing profiles in that we separate the *annotation files* from the *annotation schemata* used within them (the latter is discussed in section 3.1.). The description of annotation files themselves is limited to technical and organizational metadata. Each annotation file component is linked to the corresponding MultimodalAnnotation component, this way information about an annotation system only needs to be stored once in each session CMDI file.

3.3. Links between components

In order to better reflect the structure of the corpora, many components can be linked to each other using *attributes*. Components, which can be linked to, possess an 'ID' attribute, components which can link to other components possess 'reference' attributes. The component Actor, for instance, has an attribute ActorID, which can be linked to from components such as ActorRelation and MultimodalAnnotationFile through their ActorRef attributes.

4. Corpus metadata

The MultimodalSessionProfile is designed to describe a single set of contiguous data, usually one recording session as part of a larger corpus. For the description of the corpus as a whole, another profile is needed, 'framing' the session data. Therefore, we extended the media-corpus-profile by BAS with components

for multimodal data. The first version of the profile was mostly geared towards speech corpora containing audio data. The extended version (v1.1)[11] additionally contains a `MultimodalCorpus` component capturing information about modalities and an `AnnotationInfo` component with information about the annotated phenomena and the annotation tools and file formats used. Both the `MultimodalSessionProfile` and the extended version of the `media-corpus-profile` are meant to be used together in order to create a complete corpus metadata description.

5. Discussion and outlook

In this paper, we presented a metadata profile aimed specifically at the needs of researchers working on multimodal communication data, which builds upon and expands earlier profiles. The presentation of first approaches and the realisation of the metadata profiles have evoked fruitful discussions at conferences and workshops, both within the CLARIN community and in the relevant research communities. This shows a serious interest in the topic among potential users.

The development of our metadata profile has been driven by the requirements which resulted from the work with our specific corpora. Nevertheless, we aimed at developing a flexible profile universally applicable to multimodal data, in line with the philosophy behind CMDI: "The CMDI infrastructure encourages reuse of resources [...]. Therefore, metadata that are useful to any researcher [...] is especially valuable and should be focused on first." (de Vriend et al., 2013, 1320) Assessing the ability of our profile to actually generalize beyond our three corpora will require tests involving corpora from other research groups.

To date, user-friendly tools for the creation of CMDI files based on the `MultimodalSessionProfile` are not yet available. This remains a challenge, as the profile is too complex to create larger numbers of CMDI files by hand (especially considering the possible links between many components). Technical metadata can be easily extracted automatically from the data itself, but for the content metadata, easy-to-use tools for researchers are required and remain future work.

6. Acknowledgments

We thank Florian Schiel, Menzo Windhouwer and Onno Crasborn for their support and cooperation of the multimodal corpus profile. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication", the Center of Excellence 277 "Cognitive Interaction Technology" (CITEC), and CLARIN-D, the German division of the "Common Language Resources and Technology Infrastructure".

7. References

Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T., and Trippel, T. (2012). CMDI: a Component Metadata Infrastructure. In *Proceedings of the workshop "Describing LRs with Metadata", LREC 2012*.

Crasborn, O. and Hanke, T. (2003a). Additions to the IMDI metadata set for sign language corpora. Agreements at an ECHO workshop, May 8–9, 2003, Radboud University, Nijmegen. `http://www.ru.nl/publish/pages/522090/signmetadata_oct2003.pdf`.

Crasborn, O. and Hanke, T. (2003b). Metadata for sign language corpora. Background document for an ECHO workshop, May 8–9, 2003, Radboud University, Nijmegen. `http://sign-lang.ruhosting.nl/echo/docs/ECHO_Metadata_SL.pdf`.

Crasborn, O. and Windhouwer, M. (2012). ISOcat data categories for signed language resources. In Efthimiou, E., Kouroupetroglou, G., and Fotinea, S.-E., editors, *Gestures in embodied communication and human-computer interaction*, pages 118–128. Springer.

de Vriend, F., Broeder, D., Depoorter, G., van Eerten, L., and van Uytvanck, D. (2013). Creating & testing CLARIN metadata components. *Language Resources and Evaluation*, 47(4):1315–1326.

Freigang, F. and Bergmann, K. (2013). Towards metadata descriptions for multimodal corpora of natural communication data. In *Proceedings of the workshop "Multimodal Corpora: Beyond Audio and Video", IVA 2013*.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The META-SHARE metadata schema for the description of language resources. In *Proceedings of LREC 2012*.

Hassemer, J. (2014). *Towards a Theory of Gesture Form Analysis: Principles of gesture conceptualisation, with empirical support from motion-capture data.* Ph.D. thesis, RWTH Aachen University.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2013). Data-based Analysis of Speech and Gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its Applications. *Journal on Multimodal User Interfaces*, 7(1–2):5–18.

Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, A.-L., Braffort, A., Glauert, J., and Safar, E. (2012). Dicta-Sign – Building a Multilingual Sign Language Corpus. In Crasborn, O., Efthimiou, E., Hanke, T., Kristoffersen, J., and Mesch, J., editors, *LREC Workshop Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 117–122.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press, Chicago and London.

[11] `http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1387365569699`