First Steps towards a Tool Chain for Automatic Processing of Multimodal Corpora

Peter Menke[†], **Farina Freigang**[*]
Thomas Kronenberg[*], **Sören Klett**[*], **Kirsten Bergmann**[*†]

[*]Faculty of Technology, Center of Excellence "Cognitive Interaction Technology" (CITEC)
[†]Collaborative Research Center "Alignment in Communication" (SFB 673)
Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany
pmenke@techfak.uni-bielefeld.de, farina.freigang@uni-bielefeld.de

Abstract

This article presents the conception of a tool chain system that is capable of performing various operations on data sets from multimodal interaction research. We argue that such a tool chain can improve the workflow of scientists by providing a flexible way of performing recurring tasks from the area of creating, postprocessing, and analysing multimodal data collections.

1 Introduction

In the last decades, various scientific disciplines directed their attention to the analysis and explanation of multimodal interactions. Among other factors, two major developments fostered this change of direction: The advances and reductions of expenses of modern technology, and the trend towards interdisciplinary research with several emerging subdisciplines.

Technology. In the middle of the 20th century, the analysis of human motion was a cumbersome process. Take the example of film analysis described in Efron (1972). Here, movie projectors were used to screen the recorded body movements frame by frame onto a paper chart. Researchers then added markings and drawings to this chart that represented the interesting properties of body movements. Today, in contrast, such processes can be performed in fractions of the time with the aid of digital technology and specialised software.

Interdisciplinarity. Today, several interdisciplinary strands of research emerged that benefit from the scientific traditions of each of their ancestors. For instance, linguistics on its own dealt with an abstract, idealised concept of speech for a long time. As soon as subdisciplines emerged which engaged in an interplay with other areas of research, a shift in focus could be observed. Psycholinguistics, cognitive linguistics, sociolinguistics, and many other directions combined the expertise in language and speech with theories, methods, and scopes from the neighbouring disciplines in order to obtain a more complete picture of speech. Several of these interdisciplinary approaches made advances to the fact that spoken language is only one of many different ways that can be used to communicate and interact. Communication between humans takes part mainly on the acoustic and visual channels. Thus, it is not surprising that video and audio recordings are two fundamental ways of preserving volatile interactions and, thus, making them available as material for research (recall the situation used by Efron).

The production of such recordings, and their subsequent analysis and transformation into so-called multimodal corpora is still quite costly. One of the main reasons for this is that video and audio recordings *per se* are often not suited for a direct analysis. They have to be transformed into representations on which analyses and measurements can be performed. Such transformations, in turn, still have to be performed by human transcribers and annotators in most cases.

However, research on multimodal interactions has two properties that have consequences for the workflow of corpus assemblage and analysis: First, such research endeavours are often novel, innovative, and, thus, do not (and cannot) follow established routines and use canonical software. Second, since research on multimodal interaction happens in several different (and not necessarily adjacent) scientific fields, the approaches, methods, tools, and operations used for corpus creation and processing can diverge fundamentally.

Thus, the landscape of methods and tools for multimodal interaction analysis is changing, unstable, and nonuniform. One consequence is that it is often difficult to *reuse* or *combine* tools and methods if they come from different areas of this landscape. However, researchers could benefit substantially from a situation in which such a reutilisation and combination of different singular tools and operations was possible.

One example of a situation in a neighbouring area is WebLicht (Hinrichs et al., 2010), a tool chain for linguistic data represented in textual form. This tool chain provides several small tools in the shape of modules, each of which provides a singular operation on the text (such as a tokenisation, the addition of part-of-speech tags, the addition of named entity tags, etc.). These modules can then be chained into a sequence that performs a large, complex operation – provided that for each step in the chain, the output of the previous module provides information that is sufficient for the subsequent step to perform. One prototypical instance of such a chain is a sequence of operations that starts with a plain text representation of a text and delivers a tokenised, complex representation of the same text, together with part-of-speech, lemma, and phrase structure information.

Researchers working with multimodal data would clearly benefit from (semi-)automatic operations for certain tasks, and also of advanced technologies, such as tool chains that make it possible to glue together such singular tasks. However, since multimodal data is more complex and nonuniform than textual representations of language (due to the properties of the research landscape described above), the creation of such a tool chain is a challenging endeavour.

In this article, we present first thoughts and steps towards a tool chain that covers a small sample of multimodal interaction data sets and operations. The tool chain concept we propose is based on the analysis of multimodal corpora and experience with existing tools. Major components of our concept are an adequate data exchange format, interfaces to (semi-)automatic annotation tools, the integration of statistical analysis services, and an approach to verify the plausibility of annotations via behavior simulation with a virtual human. A first prototype implementation of this multimodal tool chain has been implemented and will be discussed with respect to benefits and limitations of tool chains for multimodal data.

This work has been initiated by the CLARIN-D discipline specific working group on "Speech and Other Modalities", which is concerned with integrating multimodal resources and tools into the CLARIN infrastructure. While the first curation project successfully *integrated* three different (gesture and sign language) corpora, the second project explored explored theoretical concepts and implementation possibilities of a tool chain for multimodal data, which are discussed in this abstract. Due to the short duration of this curation project, we did not aim at a full implementation of a runnable prototype of such a tool chain. Instead, we focused on aspects of conceptualization and feasability, and we implemented single elements of that tool chain as proofs of concepts. The description and documentation of our work in this article is supposed to serve as an aid to other researchers who are concerned with similar projects.

It is important to us to make clear that our approach does not aim at a complete and exhaustive modelling of all possible flavors and variations of human interaction. Our objects of research is not human interaction itself, and our goals, theories and hypotheses do not attempt to explain, predict, or compare such interactions. Instead, we operate on a meta level and focus on those kinds of representations, formats, and structures which we found to be typically and commonly used for the modelling of instances of human interaction. Therefore, the tool chain we propose in this work explicitly does not claim to be representative of or sufficient for a description of such interactions. It has merely been designed to be suitable for the processing of that kinds of data structures that have been used by others for research in the area of human interaction.

2 Motivation for a tool chain for multimodal interaction data

2.1 Multimodal data

First, we provide some clarification of what we consider multimodal data, since terms from the area of modalities and multimodality are not used uniformly in literature (Bonacchi and Karpiński, 2014). In order to achieve this, we present some examples of data collections that, according to our definitions, are multimodal corpora. They have different characteristics and components and thereby illustrate the spectrum and variety of possible data sets which the tool chain aims to address.

We consider data sets multimodal that contain representations of interactions that incorporate more than one functional modality. As a functional modality we define a way of communication that is characterized by the use of a distinct *coding* or *representation system* and the choice of a *channel*, consisting of a characteristic pair of emitters (or production systems) and perception systems. As an example, facial expressions are considered a distinct functional modality because the components of the face of the sender and the visual sensory pathway of the receiver[1] form a pair that instantiates a channel, and the different combinations of movements of components of the face are considered a system of representations that can, among other things, be used for the expression and communication of emotional states (as described in, among others, Ekman and Friesen 1978, Smith and Scott 1997, or Kappas et al. 2013).

As mentioned in the introduction, since interactions themselves are volatile and transient *events*, and, therefore, cannot be considered actual data (cf. Lehmann, 2005, p. 180), researchers need to operate on material that can serve as adequate representations of these events. There are two large groups of data sets that are important here: signal data, which typically result from automated processes of recording or tracking some aspects of the interaction, and transcriptions or annotations, which, in the default case, result mainly or completely from human efforts such as descriptions and interpretations of those sub-events in the interaction that are related to the research questions at hand.

While there are conventions, systems, and standards for the transcription of spoken utterances (such as the IPA symbol inventory for phonetic transcriptions), several approaches to the transcription, annotation, and representation of paraverbal and nonverbal events have been proposed and used. We cannot provide an exhaustive summary and comparison of all formalisms, theories and vocabularies at this place. Nevertheless, these systems all have in common that they provide a mapping from classes or types of events to an inventory of written or drawn, sketched, or painted symbols. The process of segmenting, classifying and categorising the interaction under observation into these distinct events must (in the normal case) be performed by human annotators. For instance, the HamNoSys notation system for sign languages (Hanke, 2004) is applied by first identifying several components of signs (such as the handshape, the orientation, or the location) and then coding these components using the appropriate graphical notations from the HamNoSys inventory.

Similarly, facial expressions can be described using the Facial Action Coding System (FACS; cf. Ekman and Friesen 1978). Here, a segmentation by components of the face (so-called *action units*) is required before each of these components can be represented with one or more FACS symbols.

While some of these systems have become rather common and widely used, researchers still regularly have to resort to the adaptation of existing solutions to their needs, or even to the creation of entirely new coding schemes. One of the main reasons for this is that research questions in the area of multimodality address recently identified, scarcely investigated phenomena for which it is not guaranteed that existing coding schemes are suited for their description. As a consequence, we briefly present some data sets that incorporate such novel or idiosyncratic representations of interactions.

[1] Although we use simple mechanistic terminology here, we do not claim that human interaction and communication can be reduced to such simple mechanisms. However, since we are concerned mainly with data modelling, we attempt to be be as theory-agnostic as possible here. Therefore, we chose such mechanistic and abstract terms.

2.1.1 The Bielefeld Speech and Gesture Alignment (SaGA) Corpus

The corpus (Lücking et al., 2013) consists of 25 dyads which are engaged in a gesture-eliciting spatial communication task combining direction-giving and sight descriptions. The scenario invoked participants to communicate information about the shape of objects and spatial relations between them. Primary data consists of audio and video data from the dialogues. The data has been systematically annotated based on annotation manuals that have been developed according to theoretical considerations and refined in pilot annotation sessions. Annotation layers divide naturally into two different partitions (cf. figure 1): *Speech annotations* comprise a transcription of spoken words, part-of-speech tags, syntax annotation and annotation of the dialogue context. *Gesture annotations* comprise a segmentation of gesture phases, coding of gestural representation techniques as well as physical gesture form annotations in terms of handedness, handshape, hand position, palm and finger orientation, and movement features. One design criterion of the SaGA corpus was the separation of a morphological description of gestures from annotations that are concerned with their semantic, semiotic and pragmatic interpretation. In order to achieve this, the corpus creators assembled their vocabularies from a variety of sources, most of them seminal works in the area of gesture research (among others, Efron 1972; Kendon 1972, 1980, 1988; McNeill 1992, 2005 or Müller 1998).

2.1.2 The Chat Game Corpus

This data set (Menke et al., 2013) is a collection of computer-mediated interactions implementing an object arrangement paradigm: One participant is presented with a target configuration of a set of colored shapes on a screen, which he has to describe via chat messages to the other participant, who has to recreate the configuration based on this information.

Time-stamped chat messages and object movements (modeled as two-dimensional coordinates) are logged using a custom XML-based file format. In post-processing, an orthographic normalization and a subsequent automated syntactic analysis – using the Stanford Parser for German (Rafferty and Manning, 2008) – was performed, and sentence types (e. g., imperative vs. indicative) and politeness markers were annotated.

Based on this data, researchers investigated the influence of description strategies, syntactic patterns, and politeness on efficiency and task success. This corpus is considered to be multimodal because it contains representations of written utterances as well as representations of actions, which are single events during the course of a larger interaction. These actions can serve different communicative functions, and they can also be the subject of further interaction and communication. Therefore, we included them as a second functional modality.

While these two corpora certainly are not representative of the immense variety of data sets based on multimodal interactions, they serve the purpose of illustrating some issues we identified when working with such corpora. These issues are described and analyzed in the following subsection.

2.2 Problems with multimodal data processing

First and foremost, in order to process corpora like the ones mentioned, established transcription and annotation tools are employed such as Praat (Boersma and Weenink, 2001), ELAN (Wittenburg et al., 2006), Anvil (Kipp, 2001), iLex (Hanke et al., 2010), or EXMARaLDA (Schmidt, 2002). While those are powerful tools with a wide range of possible operations, there are still several issues when working with them:

2.2.1 Data structures

The aforementioned tools are limited with respect to the data structures they support (Wittenburg, 2008). This is mostly due to the developer's assumption of a certain conceptual or methodological approach in data generation. Especially, one assumption made by the developers that does not always hold is that macrostructures, such as utterances or sentences, are generally annotated previous to microstructures, such as subordinate words, morphemes, or phonemes (Wittenburg, 2008, 675f.). In addition, the data of most corpora is

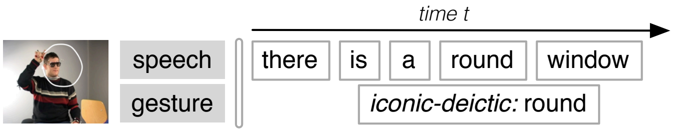

Figure 1: An annotation example of speech and gesture on two different tiers from the SaGA corpus.

time aligned, meaning that media and annotation files refer to the same time stamps. However, multimodal data is very diverse and other referent structures exist, such as spatial information, e g. about the *locations* of critical objects in the Chat Game Corpus.

2.2.2 Merging

Since each annotation tool has its specific strengths and weaknesses (Schmidt et al., 2009), users sometimes are forced to use multiple of these tools for the generation of different subsets of data. However, the merging of the resulting data is difficult or even impossible, mainly due to incompatible data structures. While many tools have import and export routines to other formats, these still are limited by the constraints of both formats. Researchers often overcome these limitations by writing scripts and similar software that combine distributed annotations of the same events, but such an approach is rather inefficient. Especially we assume that often, transformation and conversion routines are developed and implemented again and again by different researchers who did not succeed in identifying and detecting already existing solutions for their tasks.

2.2.3 Postprocessing

While each transcription and annotation tool provides operations for search, analysis, visualization, and export, in more complex cases these facilities can be stretched to their limits (Rohlfing et al., 2006). In such cases, researchers, again, regularly create ad-hoc solutions, such as scripts or macros in spreadsheet software. Often, a lot of time and manpower is spent for basic steps such as parsing the proprietary file format of the tool, finding certain patterns, and normalizing labels (such as layer names, values from controlled vocabularies, etc.).

2.3 Requirements for a tool chain

Based on these characteristics of both multimodal data sets and typical processes working with them, we consider the following requirements for a tool chain for multimodal data sets important.

2.3.1 Improved data handling via a single exchange format

The tool chain should provide an efficient way of handling proprietary data formats. We consider the use of *a single exchange format* the best choice, since the integration of new tools then only has to take into account a single file format, instead of *n* implementations for all supported formats. WebLicht uses its own exchange format, but this format has been devised for flat, text-based corpora, which renders it unsuitable for multidimensional data streams, as they are often present in multimodal corpora. For instance, parallel speaking by multiple participants, maybe also accompanied with annotations for gestures, gaze, or facial expressions, cannot be condensed into a single, flat stream of tokens without discarding vital pieces of information (cf. Figure 1).

Such an exchange format should be capable of expressing all data present in proprietary formats, and a facility for merging documents should exist. Also, the format should explicitly model data structures and content types in order to facilitate the application of tools (that is, it should be able to determine whether a document is suited as input for a certain tool).

2.3.2 Integration of existing tools

Another requirement concerns the tools themselves. Since the wheel should not be reinvented, the tool chain should be able to include already existing implementations and services. Of relevance are, for example, many tools available in WebLicht (e. g., for adding part-of-speech information to a stream of words). However, as we already stated above, the WebLicht tool chain itself is difficult to use due to the fundamental differences in the file format it uses for exchange.

Other tools are implementations for statistical analyses, gesture or sign segmentation, and the generation of movement playback in a virtual humans. Thus, the tool chain implementation becomes a collection of differently motivated services integrated in one larger infrastructure.

2.3.3 Modularisation and specification

When devising tools for the integration into a tool chain infrastructure, it should be taken care of selecting an appropriate level of granularity. Tools should act as modules that serve one single, not too complex task. Also, formal specifications are needed for the input to such modules, and for the output it produces. It must be possible to determine whether a tool is capable of handling a certain data set just by its specification, not by testing whether it succeeds or fails.

3 An architecture for a tool chain for multimodal data

Based on the specified requirements, we (1) specified a pilot instance of a tool chain that is able to handle data typical to multimodal interaction research, and (2) we implemented selected parts of it. We started with concepts from text and speech corpora and complement them with tools suited for multimodal data sets. In the following, we report theoretical considerations and the realized tool chain for multimodal data.

3.1 Theoretical considerations

In our conception of the tool chain, we made the following decisions for the handling of the requirements described above.

3.1.1 Exchange format

We decided to establish and use a single exchange format for the chosen proprietary formats of the annotation tools Praat, EAF, Anvil, iLEX, and EXMARaLDA (and also for plain text).[2] This reduces the number of implementations for adding a new functionality. Because many other details depend on this, the selection of an exchange format was the first step. Several formats were available, such as the proposal for an exchange format in Schmidt et al. (2009), BML (Behavior Markup Language; Kopp et al. 2006), and the WebLichtText Corpus Format (TCF; Heid et al. 2010). However, only the *Format for Extensive SpatioTemporal Annotation* (FiESTA) format (Menke et al., 2013) could guarantee a lossless exchange between the source files and the exchange format.[3]

[2] One reason against commercial providers is that those are more difficult to integrate into a non-commercial tool chain from a legal point of view.

[3] The W3C recommendation "Extensible MultiModal Annotation markup language"(EMMA; `http://www.w3.org/TR/emma/`), which appeared to be promising at first sight, had to be rejected because it serves a rather different purpose since it is "an XML markup language for containing and annotating the interpretation of user input", thus, it is concerned with computer-mediated data transfer and not with complex face-to-face interactions.

```
<FiestaDocument>
  <Head>
    <Section key="lifecycle"/>
    ...
  </Head>
  <ScaleSet>
    <Scale id="t" name="timeline" unit="s"/>
  </ScaleSet>
  <LayerSet>
    <Layer id="words" name="words">
      <PropertyMap/>
    </Layer>
  </LayerSet>
  <ItemSet>
    <Item id="i1">
      <Links>
        <LayerLink target="words"/>
        <IntervalLink min="10" max="11" target="t"/>
      </Links>
  <Data>
    <String>hello</String>
      </Data>
    </Item>
  </ItemSet>
</FiestaDocument>
```

Figure 2: Example of a FiESTA document, representing a single annotation which models the word “hello”, uttered in the temporal interval $[10.0, 11.0]$.

FiESTA uses a domain-agnostic format allowing for multiple timelines and spatial axes. Its information model is a superset of all annotation formats mentioned above, while a type and constraint system keeps track of data compatibility with these formats. Figure 2 shows a minimal example of a FiESTA annotation document. The format consists of a header containing metadata and a body including the contents (such as annotations). The fundamental unit is called an item, which can be linked to various references such as scales (e. g., space and time coordinates), layers (annotation level) and depending on the item type, sometimes also data blocks, which contains the annotated value.

3.1.2 Task selection

In order for the tool chain to be relevant, the performable tasks should meet the requirements of researchers in the community and they should. To that effect, we constructed a wish list within the working group and discussed realization possibilities of the various options. According to the results of these activities, the following tools should be core components of the tool chain:

1. **Annotation.** For (semi-)automatic data coding, we prioritize the widely-used text annotation tool WebLicht and one of the first gesture and sign segmentation tools AUVIS[4], which is an updated version of AVATecH toolkit (Lenkiewicz et al., 2013).

2. **Analysis.** For statistical analyses, *GNU R* seems promising as one of the most popular tools in the research community, which can easily be applied to multimodal data as well.

3. **Post-processing.** The working group explicitly stated the desire to play back motion capture data in virtual humans, as the anonymity of the participants is factual *and* gesture and sign language relevant features are given, as compared to in video data.

[4] AUVIS: http://tla.mpi.nl/projects_info/auvis/

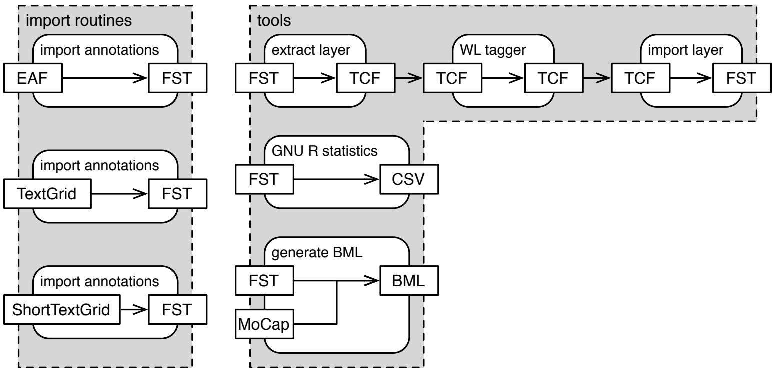

Figure 3: Modules of the multimodal tool chain. The left-hand side shows the import routines that have been implemented. The importers for TextGrid and ShortTextGrid share a common core, because both file formats are rather similar – they differ only in the surface structure of the contents of their files. The right-hand side shows the tools that can be accessed at a later stage.

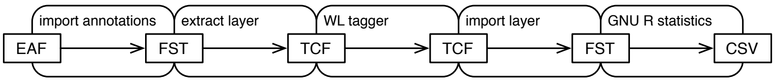

Figure 4: An instance of the tool chain that imports an EAF document, performs part-of-speech tagging in WebLicht and runs GNU R on the results.

3.1.3 Modularity

Following the tool chain paradigm, the tasks were broken down into independent modules in order to facilitate high flexibility (cf. Figure 3). They were defined by the required input format and the resulting output format. The output of a module should be reusable as input to any module that matches the resulting format. The exchange format has to support such a stepwise accumulation of data resulting from single modules. On the left-hand side of Figure 3, modules are shown that facilitate the import of third-party file formats (e. g., from annotation tools) into the FiESTA format (abbreviated as FST). On the right hand side, three exemplary steps in the tool chain are depicted.

3.1.4 Usability

Users without much technical experience should be able to use the tool chain in a simple and straightforward way. Thus, a user interface should help the user operating the tool chain (e. g., starting processes, overview ongoing processes and exporting results). However, due to the complexity of such a project, our pilot does not include any front end. However, we kept in mind the possibility of an easy attachment of such a front end.

With this conception of a tool chain system, it is possible to automatise several tasks relevant to research on multimodal interaction. One such instance of a tool chain is shown in Figure 4. Note that the graphical elements that depict input and output formats are condensed into single elements which now glue together the tool chain. This instance of a tool chain would be able to take an ELAN document (EAF is the file name extension of such files), extract a certain tier or layer, add part-of-speech information to it, reintegrate the result into the annotation document and extract quantitative data in the shape of a file with comma-separated values which is suitable to be imported into (among others) GNU R for the performance of a statistical description or analysis.

3.2 Implemented components of the tool chain

From the conception described above, a subset of modules and steps has been conducted to realize a tool chain according to the identified requirements. For each module, a RESTful HTTP service (cf. Fielding, 2000) has been implemented that facilitates the task associated with the module. Input data (in most cases, a FiESTA document or an annotation document in a third-party format) is sent via an HTTP request to the service, which is as parameterless as possible (being identified by its URL alone). The service delivers the result (again, normally a FiESTA document) in the response body of this request. This approach uses as few presuppositions as possible: Its only requirements is that the calling service is capable of performing standard HTTP requests, and that it follows the rules of the tool chain mechanism (e. g., by using FiESTA and by providing the correct kind of information in the input document).

The initial source document of a tool chain instance is an annotation file of audio and video data and at this stage of implementation, the tool chain can process Praat (TextGrid) and ELAN (EAF) files as the primary source of information. The contents of such a file is converted into the FiESTA exchange format. From here, the following services are available, implemented as three main components of the tool chain:

1. **WebLicht annotation.** The data can be processed on an optional loop through the WebLicht services for text annotation. Due to the flat primary structure of the TCF format used by WebLicht, only one annotation layer can be processed at a time. If a source file contains many layers or tiers, each is processed and stored in FiESTA separately until all desired layers were handled. In order to use the WebLicht annotation services, the data needs to be segmented into tokens and sentences. We ruled out WebLicht tokenizers for this process, since WebLicht would select different segments than the ones present in annotations, and thus, integration of the WebLicht annotations back into FiESTA could not be carried out correctly. Instead, an independent tokenizer has been implemented based on the heuristics of the WebLicht tools. This way, for each token in a TCF document, its mapping to its originating FiESTA annotation is documented. Thus, WebLicht results can reliably be mapped back to the original annotations upon reintegration into the FiESTA document. With these two auxiliary steps that map between FiESTA and a suitable TCF document, the actual text annotations can be carried out. So far, the tool chain incorporates the annotation of lemmata and part-of-speech tags.

2. **The GNU R toolkit analysis.** GNU R can be used for statistical analysis, such as the assembly of descriptive statistics or for the calculation of metrics related to interrater agreement. In this tool chain, various kinds of statistical analyses are implemented: frequency distributions, length distributions and rater-reliability with Cohen's kappa coefficient (for either different layers or different files).

3. **Behavior Markup Language (BML) file generation.** A BML file can be generated from hand shape (which can be extracted from annotations) and wrist location data, describing the movement and shape of a hand at a certain point in time. This information is merged and enhanced with specific details of the virtual human and packed into an BML file, which can then be used by a special piece of software to have a virtual character reenact the gestures represented in the file. This is only one, simple use case of BML that is supposed to demonstrate that it is possible in principle to generate usable BML data.

3.3 Discussion and evaluation

With the single elements described above, we have a very early proof of concept for a tool chain. This proof of concept was run by manually calling the items of the tool chain using the networking tool `curl` on the command line. More specifically, we created scripts that performed the necessary sequence of HTTP requests to the different services using `curl`. Up to now, the pilot implementation does not contain a system for orchestration (that is, for the automatic calling of services in the right order and with the correct stage of documents). Such an orchestration would be part of the implementation of the front end, which is planned to be carried out at a later stage, as mentioned above. As a consequence, the calls to the services have to be controlled by the user.

The proof of concept was evaluated in a very rudimentary way by using snippets from the SaGA corpus as input, and by observing that each step in the tool chain succeeded. Given the simplicity of the pilot implementation, a more complex evaluation was not feasible.

However, there are several aspects that, in our estimation, need to be covered as soon as a more complex implementation of a tool chain is present.

Runtime and sychronicity. The runtime or response time of the single tool chain steps is crucial. Multimodal corpora can become large in comparison to plain-text based ones, with the consequence that it could become necessary to transfer large amounts of data between the client and the tool endpoints (for instance, when working with complex tracking data). While WebLicht makes use of synchronous requests (where the tool orchestrator waits for each step to be finished), this could become a problem for such large amounts of data. Instead, it could be more suitable to make use of asynchronous calls to the tools (for instance, if a processing time of several hours can be expected).

Privacy. It is often the case that due to privacy issues, the recordings of participants must not be shared with a larger community. Here, it is problematic to send sensitive data (that is, data that might contain personal information) to tool endpoints.

4 Future work

Due to the restricted amount of time and working power in our project, it was possible only to create a small subset of a first working prototype. Thus, there are many possibilities of how to extend this tool chain. In the following section, ideas and options of the already connected services are sketched as well as new components and further structures.

Enhancement of integrated tools. Already integrated tools could be extended with further options. For example, the FiESTA exchange format is planned to be extended by integrating more annotation formats such as Anvil, iLex, and EXMARaLDA. Also, more use cases that take into account the possibilities of BML could be designed.

User interface. A user interface and orchestrator should be developed as a front end in order to make this tool chain attractive for users who are less proficient in scripting and programming. The interface should include job monitoring and provide first as well as partial results.

Movement segmentation. One service which is desirable to use with multimodal data in the future is the AUVIS (or AvaTech) infrastructure for linguistic processing of multimodal data. Since systems for the automatic detection of gestures have been published only quite recently, the integration of this technology into a multimodal tool chain is of major importance. A cooperation with the Auvis project has been initialized and further collaboration is highly desirable.

Metadata generation. Other considerations are to integrate results from the first curation project into a future version of this tool chain: building a service that supports the metadata generation for multimodal data. A corresponding metadata profile (Freigang et al., 2014) has been developed.

5 Conclusion

In this paper, we presented a draft for a tool chain suited to the needs of researchers working with multimodal corpora. Although only a part of the tool chain has been be implemented so far, we consider this a relevant contribution to the multimodal interaction community:

First, once a first working version of such a tool chain is available, researchers can overcome many problems occurring with multimodal corpora. Secondly, this endeavor can help to integrate the field of multimodal interaction studies into linguistic data infrastructures, where at the moment it is underrepresented compared to other branches of linguistics (such as those dealing with classical corpora based on plain text).

We are also convinced that, with the documented modular approach of such a tool chain, it is easier for others to create interfaces that add their existing tools and make them accessible from the tool chain.

Acknowledgements

We are grateful to the reviewers for their helpful comments and observations. Also, we thank the members of the CLARIN-D working group 6 “Speech and Other Modalities” who participated actively in developing the tool chain architecture and formulating most of the requirements. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) *(1)* in the CRC 673 “Alignment in Communication” (in the projects B1 “Speech-Gesture Alignment” and X1 “Multimodal Alignment Corpora”), and *(2)* in the Center of Excellence 277 “Cognitive Interaction Technology” (CITEC), and *(3)* by CLARIN-D, the German division of the “Common Language Resources and Technology Infrastructure”.

References

Boersma, Paul, and David Weenink. “Praat, a system for doing phonetics by computer.” *Glot International* 5, 9/10: (2001) 341–345.

Bonacchi, Silvia, and Maciej Karpiński. “Remarks about the use of the term "multimodality". A Word from the Editors of the Journal.” *Journal of Multimodal Communication Studies* , 1: (2014) 1–7.

Efron, David. *Gesture and Environment*. New York: King’s Crown Press, 1941.

———. *Gesture, race and culture*. Reprint of Efron (1941). The Hague: Mouton, 1972.

Ekman, Paul, and Wallace Friesen. “Facial Action Coding System: A technique for the measurement of facial movements.” *Consulting Psychologist* 2.

Fielding, Roy Thomas. *Architectural styles and the design of network-based software architectures*. Dissertation, University of California, Irvine, 2000.

Freigang, Farina, Matthias A. Priesters, Rie Nishio, and Kirsten Bergmann. “Your Data at the Center of Attention: A Metadata Session Profile for Multimodal Corpora.” Proceedings of the CLARIN Annual Conference 2014, 2014.

Hanke, Thomas. “HamNoSys. Representing Sign Language Data in Language Resources and Language Processing Contexts.” In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal, 2004.

Hanke, Thomas, Lutz König, Sven Wagner, and Silke Matthes. “DGS Corpus & Dicta-Sign: The Hamburg Studio Setup.” In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta*. 2010, 106–110.

Heid, Ulrich, Helmut Schmid, Kerstin Eckart, and Erhard W Hinrichs. “A Corpus Representation Format for Linguistic Web Services: The D-SPIN Text Corpus Format and its Relationship with ISO Standards.” In *LREC*. 2010.

Hinrichs, Erhard W, Marie Hinrichs, and Thomas Zastrow. “WebLicht: Web-Based LRT Services for German.” In *Proceedings of the ACL 2010 System Demonstrations*. 2010, 25–29.

Kappas, Arvid, Eva Krumhuber, and Dennis Küster. “Facial behavior.” In *Nonverbal Communication*, edited by Judith A. Hall, and Mark L. Knapp, Berlin/Boston: de Gruyter, 2013, chapter 6.

Kendon, Adam. “Some Relationships Between Body Motion and Speech.” In *Studies in Dyadic Communication*, edited by Aron Wolfe Siegman, and Benjamin Pope, New York: Pergamon, 1972, 177–210.

———. “Gesticulation and Speech. Two Aspects of the Process of Utterance.” In *The Relationship of Verbal and Nonverbal Communication*, edited by Mary Ritchie Key, Berlin/Boston: De Gruyter Mouton, 1980, 207–228.

———. “How gestures can become like words.” In *Cross-Cultural Perspectives in Nonverbal Communication*, Toronto: Hogrefe, 1988, 131–141.

Kipp, Michael. “Anvil. A generic annotation tool for multimodal dialogue.” In *Seventh European Conference on Speech Communication and Technology*. ISCA, 2001.

Kopp, Stefan, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. “Towards a common framework for multimodal generation: The behavior markup language.” In *Intelligent virtual agents*. Springer, 2006, 205–217.

Lehmann, Christian. “Data in linguistics.” *The Linguistic Review* 21, 3-4: (2005) 175–210.

Lenkiewicz, Przemyslaw, Eric Auer, and Sebastian Drude. “Semi–Automated Annotation of Recordings in the Humanities with Web–Services.” In *Poster presented at the Soeterbeeck eHumanities Workshop, Ravenstein, The Netherlands*. 2013.

Lücking, Andy, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. “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: (2013) 5–18.

McNeill, David. *Hand and mind.* Chicago: University of Chicago Press, 1992.

———. *Gesture and Thought.* Chicago: University of Chicago Press, 2005.

Menke, Peter, John McCrae, and Philipp Cimiano. “Releasing multimodal data as Linguistic Linked Open Data: An experience report.” In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*. Pisa: Association for Computational Linguistics, 2013, 44–52. `http://www.aclweb.org/anthology/W13-5507`.

Müller, Cornelia. *Redebegleitende Gesten. Kulturgeschichte, Theorie, Sprachvergleich.* Berlin: Spitz, 1998.

Rafferty, Anna N., and Christopher D. Manning. “Parsing three German treebanks: Lexicalized and unlexicalized baselines.” *Proceedings of the Workshop on Parsing German* 40–46.

Rohlfing, Katharina, Dan Loehr, Susan Duncan, A Brown, A Franklin, I Kimbara, J T Milde, F Parrill, T Rose, Thomas Schmidt, and Others. “Comparison of multimodal annotation tools.” *Gesprächsforschung* 7: (2006) 99–123.

Schmidt, Thomas. "EXMARaLDA - ein System zur Diskurstranskription auf dem Computer." *Arbeiten zur Mehrsprachigkeit, Folge B* 34: (2002) 1 ff. `http://www.exmaralda.org/files/AZM.pdf`. DE.

Schmidt, Thomas, Susan Duncan, Oliver Ehmer, Jeffrey Hoyt, Michael Kipp, Dan Loehr, Magnus Magnusson, Travis Rose, and Han Sloetjes. "An exchange format for multimodal annotations." In *Multimodal corpora*, Springer, 2009, 207–221.

Smith, C. A., and H. S. Scott. "A componential approach to the meaning of facial expressions." In *The Psychology of Facial Expression*, edited by James A. Russell, and José MIguel Fernández-Dols, Cambridge (UK): Cambridge University Press, 1997, 229–254.

Wittenburg, P, H Brugman, A Russel, A Klassmann, and H Sloetjes. "ELAN: a professional framework for multimodality research." In *Proceedings of Language Resources and Evaluation Conference (LREC)*. 2006.

Wittenburg, Peter. "Preprocessing multimodal corpora." In *Corpus Linguistics. An International Handbook*, edited by Anke Lüdeling, and Merja Kytö, Berlin: Mouton de Gruyter, 2008, volume 1, chapter 31, 664–685.