The Ariadne System: A flexible and extensible framework for the modeling and storage of experimental data in the humanities

Peter Menke, Alexander Mehler

Project X1 “Multimodal Alignment Corpora”
CRC 673 “Alignment in Communication”
Universität Bielefeld, 33615 Bielefeld, Germany
{Peter.Menke,Alexander.Mehler}@uni-bielefeld.de

Abstract

This paper introduces the Ariadne Corpus Management System. First, the underlying data model is presented which enables users to represent and process heterogeneous data sets within a single, consistent framework. Secondly, a set of automatized procedures is described that offers assistance to researchers in various data-related use cases. Finally, an approach to easy yet powerful data retrieval is introduced in form of a specialised querying language for multimodal data.

1. Introduction

During the last decades, interdisciplinarity has become a central keyword in research. As a consequence, many concepts, theories and scientific methods get in contact with each other, resulting in many different strategies and variants of acquiring, structuring, and sharing data sets. See Table 1 for an overview of relevant data types used by a large research centre investigating dialogical communication. Notice the wide range of data formats. Such a spectrum of representation systems leads to a problem: Researchers regularly need to work with multiple software tools whose data formats are incompatible. While there are solutions of data conversion for a few combinations of data formats (e.g., integration of Praat transcriptions into ELAN annotation documents), this does not hold in general. As an example, body tracking data cannot be added to such a document, since the corresponding tracking software uses a custom data format which cannot be read by either Elan or Praat. This means that there is no single software system exists that can handle both data subsets simultaneously. The lack of software to fill that gap was a fundamental motivation for the design of the software system presented in this paper: The Ariadne Corpus Management System. It has been built around a generic model of dialogical events oriented at central scales. These provide an abstract model of the widespread data spectrum observable in dialogical communication. The data model contains a rich type system that helps to put dialogical events into a well-defined conceptual grid, thus ensuring that data can be handled uniformly in every case. For the various proprietary data formats required by different user groups, porting routines have been designed that map between custom data models and the equivalent data structure in the Ariadne model.

2. Demarcation from other tools

There is a variety of corpus management systems, each of them accomplishing a slightly different purpose. However, we did not find a project or application that met all conditions that we considered important for our research centre, namely:

	MMAX2	*Annex/Imdi*	*EXMARaLDA*
Availability	local	**web-based**	local
Editing	**yes**	no	**yes**
Scope	narrow	**wide**	medium

Table 2: Key feature evaluation of some existing corpus management applications, namely MMAX2 (Müller and Strube, 2006), “Annex” and “Imdi Browser” (Wittenburg et al., 2002) and the ExmarAlda corpus manager (Schmidt, 2002; Schmidt and Wörner, 2005).

1. The software should be available instantly and at as many places as possible, so only web-based approaches were considered that did not require software installation or configuration. A plain web-browser should suffice (which is available at nearly every scientist’s desk).
2. Users should be enabled to modify and edit selected data and share their products and results with other users – with individuals as well as groups. In other words, the system should support user accounts and means of granting different privileges on selected resources to others.
3. The system should support many different data formats, by being as little restrictive as possible. The system should not be a specialist for one modality or structure. Instead, it should guarantee transfers and conversions between many different data subsets.

Table 2 gives an overview of how three of the most prominent tools meet these three requirements.

MMAX2 (Müller and Strube, 2006) is a tool whose focus is research related to anaphoric relations. It covers the annotation process as well as subsequent tasks (inter-annotator agreement, corpus assembly, etc.), but it is designed for (and restricted to) linguistic events that do not need linkage to explicit time points. In various multimodal settings, however, time and temporal

modality	types of data	tools and mechanisms used
speech	orthographic transcriptions, phonetic transcriptions, selective markup of relevant keywords, syntax annotion	ELAN (Wittenburg et al., 2006), Praat (Boersma and Weenink, 2001), Anvil (Kipp, 2001), EXMARaLDA (Schmidt, 2002); custom XML-based formats
prosody	set of prosodic features	Praat
mimics	markup of facial expressions	ELAN
gesture	type and components of gesture	ELAN
spatial behaviour	position and rotation of body limbs (in humans as well as artificial agents) and of critical objects	custom format (XML and CSV)
gaze behaviour	eye tracking data	custom format (XML and CSV)
(inter)action	annotation of action phases in restricted situations (e. g., during a map task game)	ELAN

Table 1: Overview of the types and formats of scientific data sets that are currently supported and used inside the Ariadne Corpus Management System used in the Collaborative Research Centre "Alignment in Communication".

agreement is vital.
In addition, MMAX2 runs as a normal application working on data taken from the local file system. For distributed research groups like the one mentioned in Table 1, this approach has a set of disadvantages (covering merging distributed multiple partial annotations as well as the need for corpus-wide backup, etc.).

Annex and Imdi browser (Wittenburg et al., 2002), on the other hand, are web-based tools for the exploration of single annotation documents (Annex) as well as hierarchically organised corpora (Imdi browser). In this case, the second requirement is not met, since all operations are read-only, so users are not able to submit new data or changed subsets to the system.

The EXMARaLDA corpus manager (Schmidt, 2002; Schmidt and Wörner, 2005), has advantages and disadvantages that are similar to those of MMAX2 – is is running locally and was designed with dialogue data in mind. However, there is support for temporal information in annotations, making the system more flexible with respect to the modeling of temporal information.

These are not the only tools we investigated, but they are rather prominent examples of powerful corpus management tools. They can be optimal if used for the purpose they have been developed for, but each of them lacks at least one key feature we consider vital for our purposes.
As a consequence, the Ariadne system has been developed. In the following section, its data model will be deduced from observings of typical data collection processes.

3. Data model

The overall structure of the data model is motivated by typical workflows in experimental research in the field of the humanities (cf. Figure 1, the items of the enumeration correspond to the main blocks in the figure).

1. Events in reality are transient, so a way of storing information about them is required.

2. They are recorded on media (typically, audio or video), thus forming a (possibly incomplete) image of reality.

3. Normally, next comes a dialogue transcription, or an annotation of events. Here, another process of mapping takes place: The mapping from media to their basal dimensions or *scales* (e.g., timelines or spatial coordinates). In general, clear and accurate data collection is possible only if all scales of a setting are properly identified and defined.

4. Based on this model, the process of transcribing or annotating *primary data* is merely a positioning of data chunks along these scales.

5. In case of annotation of *secondary data*, these chunks of data may not only refer to scales, but also to already defined data elements.

Ariadne's data model (see Figure 2) allows the user the definition of multiple scales whose member elements are the reference points for all primary events.
Such an *event* contains the actual information in a document, and it is mainly a collection of data elements together with a list of *links* to all other elements is is related to. There are different kinds of links, depending on the type of targets:

1. *Scale links* determine a position in the model of reality defined by these scales. These "positions" can take the form of singular points, of single intervals or of compound sets of these elements, which makes it possible to define complex structures like discontinuous events.

2. *Event links* refer to other events. These could be parents, predecessors, or co-referent elements.

3. *Layer links* define memberships of events in so-called *layers*. These layers group together elements of the same type. In software systems like Elan or Praat, data elements are visualised vertically according to their layer membership.

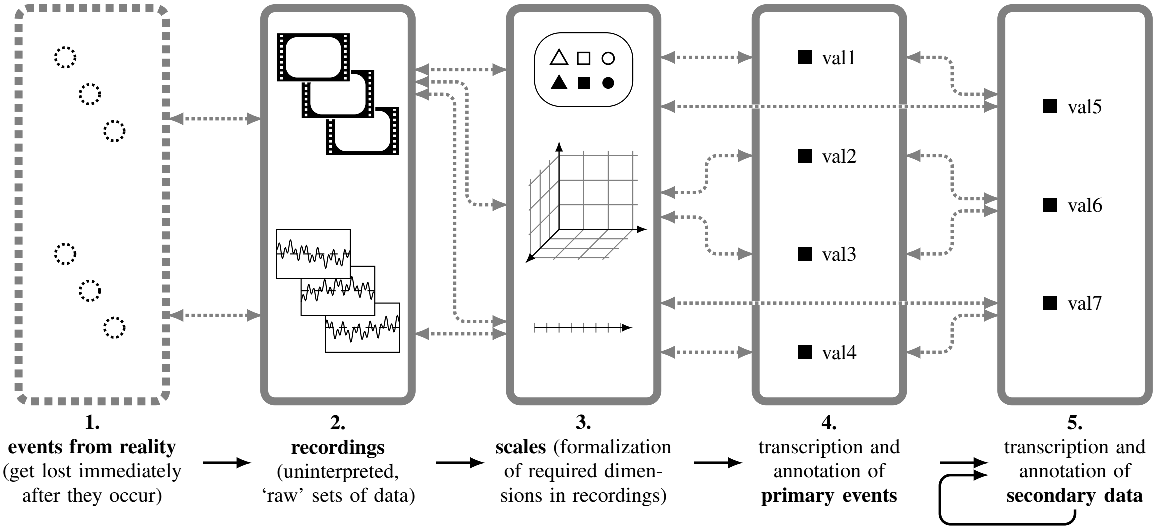
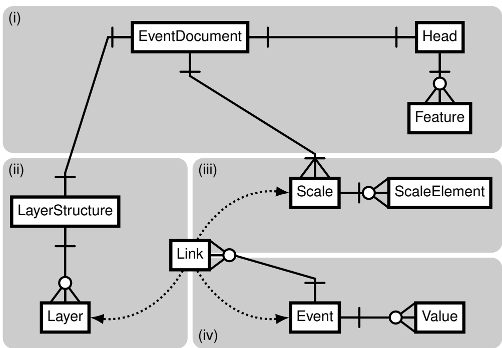
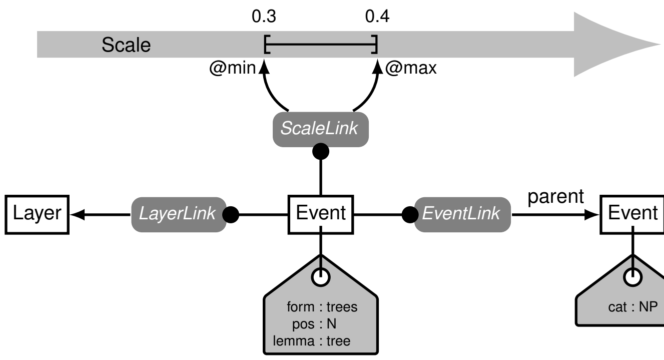


Figure 1: Workflow of different stages of data acquisition during the performance, recording, transcription and annotation of a study that typically occurs in experiments investigating dialogical communication.



(a) A simplified class diagram for an Ariadne document Central elements are events with a set of links.



(b) An event (modeling the word "trees" uttered from seconds 0.3 to 0.4) as it is represented in the Ariadne data model:

Figure 2: An structural overview of (a) a complete Ariadne annotation document, and (b) a single event, including its connections to other parts of the document. This event is linked to a time interval by a *ScaleLink* object, it belongs to a certain layer via a *LayerLink* object, it is connected to a parent event (modeling an NP) by means of an *EventLink* object, and it contains a data structure that models the event's contents (namely, information about the word's word form, its part of speech and its lemma).

This data model is sufficient for operations that go beyond human annotation of dialogical data. Several tasks in scientific workflows can be automatized, and in the following section, an apparatus for such tasks is presented in detail.

4. Data generation & enhancement

In scientific research dealing with experimental and/or corpus data, several tasks occur frequently, including

1. automatized data generation, e.g. by means of part-of-speech taggers, sense taggers, syntactic parsers;

2. the creation, application and modification of transcription and annotation schemas, in order to have resulting data match theoretical prerequisites;

3. the combination of human-generated data (like annotations or ratings) from multiple creators;

4. further processing and transformation of data (e.g., data sets as input for third-party software, or data in form of human-readable documents or as a graphical visualisation).

Ariadne provides a set of modules that have been designed for such tasks. To be concise, these modules assist users by automatising certain subtasks by providing modules and graphical user interfaces (GUIs), thus saving time and reducing the number of errors caused by humans.

4.1. Tagging and parsing of speech transcriptions

As an example of the integration of software modules into the Ariadne System, we outline the integration of the eTagger library (Gleim et al., 2009). It is a flexible part-of-speech tagger that has been trained on several German and English corpora. It has been attached to the Ariadne system, so it can be called to work on any Ariadne document that contains word form annotations. In addition, phenomena typical of spontaneous speech (hesitation signals, pauses or fragmented words) are filtered and marked during with the tagging process. The result can be loaded into

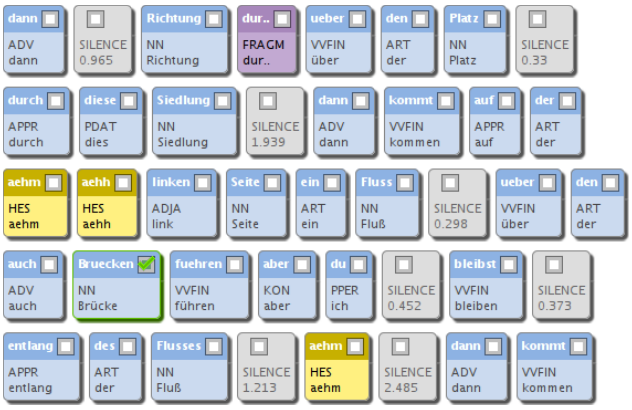


Figure 3: Extract of the visualisation of the PoS Tagger Output. Special elements are highlighted in different colors, namely silent intervals, hesitation signals or word fragments. The GUI has features for editing and correcting of single elements on the fly.

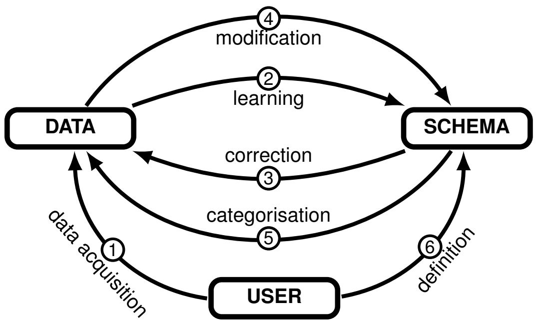


Figure 4: Exemplary use cases that deal with data sets and related data schemas.

a visualisation module of the GUI where different types of words, non-words and other elements are highlighted in different colors (see Figure 3).
Users can then interact with these elements in different ways: Faulty elements can be marked as "correct" or "incorrect", and it is also possible to select a different part of speech category and assign it to the element. These corrections are saved on the server and will be used as input for further optimization of the tagger, thus increasing its accuracy.

4.2. Schema generation & validation

Research results need to be accurate and built on top of a clear theoretical basis. Therefore, data sets must also conform to certain rules. These can be defined in different ways. To cover as many rule definitions as possible, Ariadne supports so-called schemas. These can be as simple as a set of allowed values (e.g., object names), but it can also consist of more complex parts (e.g., grammar rules). In order to provide this functionality for users, the Ariadne system provides GUIs that assist users in the following use cases (cf. Figure 4):

1. **Data acquisition:** When uploading data, users have the option to validate their data after upload and to correct or reject invalid documents.
2. **Learning of categorial information:** See 4.3..
3. **Auto-Correction of data according to a given schema:** Since data generated by humans typically contains errors (e.g., typographical errors), it can be necessary to check data sets and filter out such errors. Ariadne provides a wizard that takes a schema (for example, in the form of a set of allowed values) and a data document and produces a corrected version by comparing each value against a set of valid values. If the value is not valid, the system calculates the nearest possible member with the aid of a similarity measurement (for strings, this could be the Levenshtein distance). The value is then replaced.
   An example: Due to the lack of an explicit annotation scheme in a research project, some errors occurred in orthography and in the use of whitespace during annotation. 14.4% of all events contained errors. With the autocorrection GUI, all errors could be eliminated only by creating a list of the correct value items and handing it to the correction algorithm (cf. Figure 5).
4. **Assistance in schema modification:** Documents can be visualized in a special window where elements can be highlighted according to different rules. One possible use case is automatic markup of data that is regarded invalid according to a given schema. This can help users who want to align or redesign their schema to existing data sets.
5. **Categorisation of data according to schemas:** Simple mechanisms of data categorization can also be given in form of schemas. As an example, the markup and filtering of German hesitation signals has been defined in one central schema which (if needed) could then be applied to all documents inside the Ariadne system whose transcription approach is compatible.
6. **Explicit schema definition:** Of course, users can also import schemas directly into the system.

4.3. Export routines

There is a wide range of software that assists scientists in quantitive research, e.g., various machine learning systems, or statistical software. Since these tools often are rather extensive and complex, there is no integration or direct annexation of Ariadne to these systems. Instead, an interface for data formats has been designed. It permits a seamless and intuitive conversion of sets of dialogical data to a range of data lists that are readable by various software tools in the field of statistics or machine learning: With Ariadne, users can provide their statistical software with all data needed for computing, with just a few steps and selections in a special GUI. Again, the actual structure of input data is irrelevant as long as it can be imported into the Ariadne system.

5. Data retrieval

Finally, this section describes the prerequisites for a special query language for scale-related events as an additional part of Ariadne. Subsets of data can be retrieved matching descriptions formulated in that query language. These descriptions contain conditions regarding data values, related

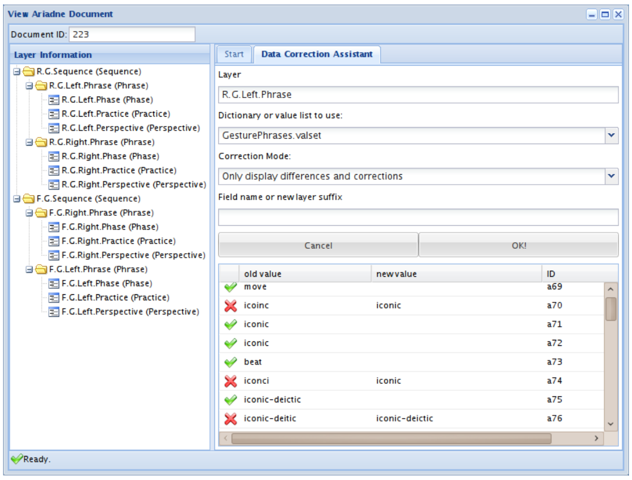


Figure 5: Correction Wizard, showing the value corrections that the correction algorithm proposes for a layer of gesture annotations.

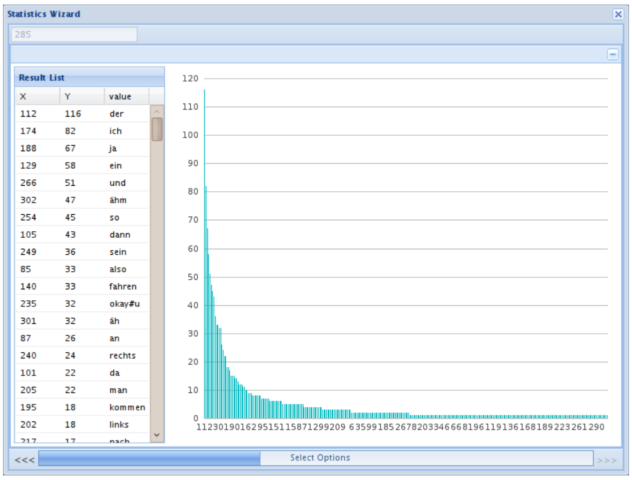


Figure 6: Statistics Wizard showing the rank-frequency distribution of lemmas occurring in one experimental dialogue.

events and scales, and may also perform operations on data elements and scale members.

While a GUI will be provided for live entry of such queries against given documents, another feature is the storage of such queries inside a document in so-called virtual layers. For a virtual layer, a set of member events can be retrieved, but as a difference to 'classical' layers, no explicit membership definition is given inside the events, but instead, members are defined as all elements that match the given query. This makes it possible to define layers whose contents is dynamic – it may change when other parts of the document are modified. Finally, virtual layer definitions can be shared as well as documents, allowing their reuse on other documents.

This module is still under construction. However, most of its fundamental functionality has already been integrated into the Ariadne data model.

6. Conclusion

With the Ariadne system, a flexible tool has been introduced that can assist users in various cases of scientific research on dialogical and multimodal data. Numerous custom data formats can be imported and combined with each other, and data can be validated against given data definitions and restrictions. Furthermore, data can be grouped, queried and reorganised with the aid of specialised queries. Finally, export routines to various third-party software tools are provided in order to simplify further data processing and analysis.

Acknowledgement Financial support of the German Research Foundation (DFG) through the the SFB 673 *Alignment in Communication* (via the Project X1 *Multimodal Alignment Corpora: Statistical Modeling and Information Management*) is gratefully acknowledged.

7. References

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

Rüdiger Gleim, Ulli Waltinger, Alexandra Ernst, Alexander Mehler, Dietmar Esch, and Tobias Feith. 2009. The ehumanities desktop - an online system for corpus management and analysis in support of computing in the humanities. In *Proceedings of the Demonstrations Session of the 12th Conference of the European Chapter of the Association for Computational Linguistics EACL 2009, 30 March – 3 April, Athens*.

M. Kipp. 2001. Anvil – a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*. ISCA.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

T. Schmidt and K. Wörner. 2005. Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung*, 6:171–195.

T. Schmidt. 2002. Gesprächstranskription auf dem Computer – das System EXMARaLDA. *Gesprächsforschung*, 3:1–23.

P. Wittenburg, W. Peters, and D. Broeder. 2002. Metadata proposals for corpora and lexica. In *Proceedings of the Third Language Resources and Evaluation Conference (LREC)*, pages 1321–1326.

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