Adapting Standard NLP Tools and Resources to the Processing of Ritual Descriptions

Nils Reiter[1] and **Oliver Hellwig**[2] and **Anand Mishra**[2] and **Irina Gossmann**[1]
and **Borayin Maitreya Larios**[2] and **Julio Rodrigues**[1] and **Britta Zeller**[1] and **Anette Frank**[1]

Abstract. In this paper we investigate the use of standard natural language processing (NLP) tools and annotation methods for processing linguistic data from ritual science. The work is embedded in an interdisciplinary project that addresses the study of the structure and variance of rituals, as investigated in ritual science, under a new perspective: by applying empirical and quantitative computational linguistic analysis techniques to ritual descriptions. We present motivation and prospects of such a computational approach to ritual structure research and sketch the overall project research plan. In particular, we motivate the choice of frame semantics as a theoretical framework for the structural analysis of rituals. We discuss the special characteristics of the textual data and especially focus on the question of how standard NLP methods, resources and tools can be adapted to the new domain.

1 INTRODUCTION

Led by the observation of similarities and variances in rituals across times and cultures, ritual scientists are discussing the existence of a "ritual grammar", an abstract underlying – and possibly universal – structure of rituals, which nevertheless is subject to variation. It is controversial whether such structures exist, and if so, whether they are culture-independent or not.

Our interdisciplinary project[3] addresses this issue in a novel empirical fashion. Using computational linguistics methods, we aim at obtaining quantitative analyses of similarities and variances in ritual descriptions, thereby offering ritual scientists new views on their data.

Ritual researchers analyze rituals as complex event sequences, involving designated participants, objects, places and times. Such sequences are usually encoded in natural language descriptions. However, the knowledge of recurrent structures in ritual event sequences is often private among researchers devoted to particular cultures or scientific fields, because an all-encompassing theoretical framework for the analysis of rituals across different cultures does not yet exist. In our work, we attempt to make characteristic properties and structures in rituals overt. For this sake, we apply formal and quantitative computational linguistic analysis techniques on textual ritual descriptions. We will investigate data-driven approaches to detect regularities and variations of rituals, based on semi-automatic semantic annotation of ritual descriptions, thereby addressing this research issue in a novel empirical fashion.

As a ritual can be divided into complex event sequences, the computational linguistic analysis of ritual descriptions needs to focus on discourse semantic aspects: the recognition and analysis of events and roles, temporal relations between events and coreference and anaphora resolution regarding participants of these events, to name just a few. In order to capture variations and similarities across rituals, it is important to analyze and quantify variations in event successions (e.g., is a specific action accompanied by another one, or strictly followed by it?), as well as variance regarding the ontological type of participants (what kinds of materials or living beings are subject to or involved in specific actions in different roles?).

Computational Linguistics resources and tools for the analysis of ritual structure. Computational Linguistics has developed a variety of resources and processing tools for semantic and discourse processing that can be put to use for such a task. The community has developed semantic lexica and processing tools for the formal analysis of events and their predicate-argument structure, in terms of semantic roles [11, 16, 24], temporal relation recognition [32], and anaphora and coreference resolution [29, 33, 23]. Using these resources and processing tools, we can compute structured and normalized semantic representations of event sequences from textual descriptions of rituals, and thus identify recurrent patterns and variations across rituals by quantitative analysis. Frame semantics [11], with its concept of scenario frames connected by frame relations and role inheritance, offers a particularly powerful framework for the modeling of complex event sequences. It can be used to structure event sequences into more abstract concepts that may subsume different kinds of initial, transitional or closing events of rituals. Through the annotation of word senses, using lexical ontologies such as WordNet [10], we can observe and analyze variations in the selectional characteristics of specific events and their roles across rituals. The integration of corpora and ontologies [2] offers possibilities to reason over corpora and external knowledge resources.

Processing ritual descriptions with standard NLP tools. The semantic annotations, however, need to be built upon linguistically preprocessed data. This preprocessing consists of several layers, starting with tokenization, part of speech tagging, and shallow or full syntactic analysis. Semantic analysis tasks, such as semantic role labeling or coreference resolution typically builds on these pre-processing levels. As a basis for semantic annotation we use existing open-source systems for tokenizing, part of speech tagging, chunking or parsing. Automatic anaphora and coreference resolution provide im-

[1] Department of Computational Linguistics, Heidelberg University, Germany
[2] South Asia Institute, Heidelberg University, Germany
[3] The project is part of a collaborative research center (Sonderforschungsbereich, SFB) on "Ritual Dynamics". Over 90 researchers from 21 scientific fields work the structure and dynamics within and across different cultures. http://www.ritualdynamik.de

portant information for a coherent textual representation based on semantic role analysis. The systems we use for this preprocessing are data-driven, and have proven to obtain high performance scores, as they are typically trained on large corpora. In fact, such statistical systems often outperform rule-based systems.

However, there is one caveat: Most currently available training (and testing) data is taken from the news domain or encyclopedias like Wikipedia, which represent one or more particular domain(s). The assumption that data-driven approaches can be applied to an arbitrary new domain relies on the availability of training data for this domain. This, however, is rarely the case, especially if we move to "small" domains featuring special linguistic phenomena combined with restricted textual sources and a complete lack of annotated textual material.

In this paper, we report on first steps to provide a proof of concept for using computational linguistic resources and analysis methods for the study of ritual structures, based on small collections of data, analyzed at all intended levels of representation. In particular, we present initial studies that assess (i) the performance of standard NLP tools and resources for processing linguistic data from the ritual domain and (ii) the need and basic methods for domain adaptation.

Section 2 presents the project research plan and related work. In section 3, we discuss special linguistic characteristics of the textual data that have an impact for automatic processing. Section 4 presents experiments that measure the performance of standard NLP processing tools on various linguistic depths and assess basic domain adaptation techniques. Section 5 presents our methodology for performing the semantic annotation of ritual descriptions by assessing the coverage of existing resources, the need for domain adaptation as well as a principled work flow to enable future automation. Finally, we present an outlook on the type of analyses we expect to produce to enable empirical studies of the structure and variance of rituals. Section 6 describes plans for future work and concludes.

2 COMPUTATIONAL LINGUISTICS FOR RITUAL STRUCTURE RESEARCH

2.1 Project research plan

The project is divided into two consecutive stages of research, which concentrate on corpus creation and annotation and on the analysis and exploitation of the data, respectively.

Corpus creation and annotation. In the first stage, a comprehensive corpus of linguistically and semantically annotated rituals from different cultures will be created from natural language descriptions of rituals that are procured by experts. The semantic annotation follows the frame semantics paradigm [11] and comprises both general linguistic and ritual-specific annotations.

As we aim at an empirical basis for the conceptualization of the domain, we automatically identify relevant domain terms on the basis of scientific publications on ritual research which in turn can serve to establish a base vocabulary for the annotation with ritual-specific concepts.

Analyzing the structure of rituals. Based on the semantic annotation of ritual descriptions, logical and statistical methods will be deployed to detect recurring structures in ritual descriptions, as well as systematic variances. In close cooperation with the ritual researchers, we will provide tools for the exploration of our data-driven, quantitative analyses of rituals.

2.2 Related work

Central to the structure of rituals are sequences of events and participants involved in these events. Hence, an important research topic is the **detection and analysis of event chains** in texts. The use of frame semantics as a useful abstraction layer for analyzing event chains has been investigated in [1]. A case study demonstrated how relations between instances of frames and roles can be inferred in context, using frame relations as well as contextual information, such as coreference or syntactic association. A related shared task on "linking roles in discourse" [27] is being organized as part of SemEval 2010. Recently, a statistical approach has been proposed for unsupervised detection of event chains, using co-occurrence of a single discourse entity as argument of different verbs as well as coreference information as criteria for extracting event chains [3, 4]. Results of applying similar linguistic and computational techniques to a corpus of Sanskrit texts are reported in [13], where chains of events are used to detect the temporal structure of a corpus.

Another central issue related to our work is **domain adaptation**, because most NLP tools are trained on news corpora. An interesting approach for addressing the domain adaptation problem is augmenting the feature space to model both domain and general, domain-independent characteristics [6]. A very similar approach employs hierarchical bayesian prior to encourage the features to take similar weights across domains, unless the differences of the data demand otherwise [12]. Both methods make use of labelled data. A contrastive approach has used an instance weighting framework, where unlabeled instances of the target domain contribute to the model estimations [15].

3 RITUAL DESCRIPTIONS

We collect ritual descriptions from different sources. The collection process has been started with Hindu rituals from Nepal and rituals from the Middle East, but we plan to extend it to rituals from Ancient Egypt and the Middle Ages in central Europe. All our methods and techniques are culture-independent and can be adapted to other, non-English, languages.

We decided to concentrate on translated ritual descriptions that have already been published in scientific literature in order to quickly collect larger amounts of data that is relevant and trustworthy. All ritual descriptions are entered by a ritual researcher. We use a trac Wiki[4] as an interface, because it (i) allows easy to follow structuring rules, (ii) is readable by every project member on the Web without knowledge of XML or other markup languages and (iii) is designed for automatic processing.

In the following, we discuss specific properties of the ritual descriptions in our corpus that are relevant from a computational linguistics point of view.

3.1 Textual sources

We use two types of textual sources. The first comprises studies by ritual researchers that deal with the religious, ethnologic and social background of rituals and are strongly theory-oriented. These texts will serve as a basis for building a ritual specific ontology, starting from a common terminology [26]. The second type of texts are descriptions of rituals. These sources form the basis of the ritual corpus and are, therefore, of special importance for the project. Two subtypes of ritual descriptions can be distinguished.

[4] `http://trac.edgewall.org`

Ethnographic observations are an important source for our knowledge of how rituals are performed in modern times. These texts are written in English, though not always by native speakers. Some scholars tend to intersperse purely descriptive passages with theoretical interpretations of what was observed. The actual course of the rituals can thus not always be separated clearly from personal interpretations (see 3.2.5).

Translations of indigenous **ritual manuals** that may date back several centuries are the second subtype of the ritual descriptions. Originally, the manuals are written in non-English languages (e.g., Sanskrit), but English translations of them have been published in ethographic literature. Contrary to the ethnographic observations, these sources are mainly prescriptive in character. Since many of these manuals are intended as a kind of memory aid for ritual practitioners, they often record only the most important or extraordinary steps of a ritual, while typical, recurrent elements are omitted. This selective choice of content complicates the alignment of such manuals with the exhaustive descriptions of modern observers.

The subtype of ritual descriptions is stored as meta data attached to the source text, along with the bibliographic source of the descriptions, original language and related types of information.

3.2 Text characteristics

Dealing with ritual descriptions requires handling of special phenomena on the lexical, syntactical and discourse-level. We describe these challenges in the following.

3.2.1 Foreign terms

A ritual description produced by a ritual expert (be it a researcher or a practitioner) often contains terminology specific to the cultural context of the ritual. In most cases, English counterparts for these terms do not exist. Therefore, they often remain untranslated in the texts (although transliterated into Latin characters).

(1) He sweeps the place for the sacrificial fire with *kuśa*.

Kuśa is a Sanskrit term for a kind of grass that is very important in Vedic rituals. For this ritual, it is necessary to sweep with *kuśa* and not any other grass.

The term *kuśa* has never been seen by a common, newspaper trained part of speech tagger nor is it contained in a lexicon of a rule-based grammar. We therefore decided to annotate such terms with English paraphrases as in Example 2. For automatic processing, the original terms are replaced by the paraphrases and are later re-inserted.

(2) He sweeps the place for the sacrificial fire with <grass * kuśa>.

3.2.2 Fixed expressions

Most rituals contain fixed expressions consisting of multiple words or sentences. These expressions are often prescribed pieces of text which have to be spoken or chanted while a ritual is performed (e.g., *Our father* in Christian church service).

(3) Salutation to Kubera reciting the mantra *arddha-māsāḥ [. . .]*;

There is no common way in handbooks or scientific literature to refer to such fixed expressions. Sometimes, prayers or chants have a title or name; sometimes, first words or the refrain are given and the expert knows the exact expression.

As most fixed expressions cannot be translated literally, we adopt them as unanalyzed expressions in a foreign language. We ask the ritual experts to mark them as such, so that we can replace them with indexed placeholders during processing and re-insert them later.

3.2.3 Imperatives

As ritual manuals are often written by and for ritual practicioners, they contain a high amount of imperative sentences. In a randomly selected sample of ritual descriptions, we found 20% of the sentences realized in an imperative construction. The ritual description with the highest amount of imperatives contains over 70% of sentences with imperative constructions. By contrast, in the British National Corpus, only about 2 % of the sentences contain imperatives.

3.2.4 PP-attachments and nested sentences

Prepositional phrases (PPs) are quite common in the data, as becomes apparent in Example 1. This introduces ambiguities that are hard to resolve. Deeply embedded PPs (4) are difficult to attach correctly, but appear in the texts regularly.

(4) . . . worship of the doors of the house of the worshipper.

The frequency of syntactic coordination and nested sentence structures is varying between languages and text types. In Sanskrit, which is the source language of most of our data, long and nested sentences are very common. This characteristic is also reflected in the texts' translations into English, as the translators (i) try to preserve the original text character as much as possible and (ii) do not aim at producing well-to-read English sentences.

The joint occurrence of PP attachment, coordinations and sentence embedding are a challenge for syntactic processing. Example 5 illustrates the interaction of coordination (*italic*) and PP attachments (underlined) in a long sentence.

(5) Beyond the members of the lineage, these visits lead to the paternal aunts of three generations which includes father's *and* grandfather's paternal aunts *and* their daughters *and* granddaughters, the maternal uncles *and* maternal aunts of their grandmother *as well as* their maternal uncles of three generations.

This leads to a combinatorial explosion of possible analyses and – in case of statistical disambiguation – a parser is deemed to make wrong guesses. Therefore, since full-fledged syntactic analyses are not necessarily needed for role semantic labeling (see e.g. [9]), we opted for flat syntactic analysis based on chunks.

3.2.5 Interpretations

Ritual descriptions that have been published in scientific literature often do not contain "clean" descriptions restricted to the ritual performance only. Instead, the description of a ritual performance is interwoven with comments or interpretations that help the reader to understand the ritual.

(6) The involvement of the nephews can be understood as a symbolic action to address those of the following generation who do not belong to the lineage of the deceased.

Example 6 is clearly not an event that occurs during the ritual, but a scientific interpretation. Although it is in principle possible to annotate such sentences with frames and frame elements, they represent a different level of information that does not belong to the ritual itself. As we want to automatically extract common event sequences from the ritual descriptions, such interpretations need to be clearly separated from descriptions of factual events.

In order to systematically address this issue, we divided the sentences into three classes:

1. Sentences that clearly indicate events happening during the ritual performance (Example 3)
2. Clear interpretations, citations or comments (Example 6)
3. Sentences that are ambiguous with respect to these classes, or sentences that contain elements of both classes (Example 7)

(7) The wife of the chief mourner [...] will carry a symbolic mat that represents the bed of the deceased [...].

We performed an annotation study on a randomly selected ritual description (40 sentences) and found that 15% of the sentences contain both interpretative and factual elements or are ambiguous (clear interpretations: 17.5%, clear factual statements: 67.5%). We did not yet experiment with automatic tagging of sentences according to their class. One possibility, however, could be the application of methods used for the automatic detection of hedges. Academic writers tend to use a high amount of hedges [14]. From the examples in our ritual descriptions, hedges indeed appear quite often. Following the definitions given in [19] and [18], 42.9% of our sentences with a clear interpretative character contain linguistic hedges. There is existing work on the automatic detection of hedges [19, 30] which may be adapted to our specific concerns.

As a first partial solution to the problem, we decided to annotate the clear interpretative sentences as such. They will be ignored for the frame annotation, but remain in the texts.

4 AUTOMATIC LINGUISTIC PROCESSING

As a basis for semantic annotation and processing, the ritual descriptions are preprocessed with standard NLP tools. We use UIMA[5] as a pipeline framework, in which we have integrated a rule-based tokenizer, the OpenNLP[6] part of speech tagger, the Stanford Lemmatizer [31] and the OpenNLP chunker.

4.1 Tokenizing

Many of our texts contain special, non-English characters (*ś*) or complete tokens (*Gaṇeśa*). Therefore, we employ a rule-based tokenizer that uses Unicode character ranges in conjunction with an abbreviation lexicon to detect common abbreviations such as *etc.* or *i.e.*

4.2 Part of speech tagging and chunking

Using standard models for part of speech tagging and chunking produces rather poor results. This is due to the fact that our data contains (i) a lot of unseen tokens and (ii) a high amount of rare and uncommon constructions. We experimented with different scenarios for the domain adaptation of an existing part of speech tagger and chunker.

[5] http://incubator.apache.org/uima/
[6] http://opennlp.sf.net

As we aim at a culture- and source language independent framework, we decided to use a statistical part of speech tagger and chunker, that can be trained on specific corpora.

Large amounts of training material for both labeling tasks are available from other domains, and the annotation of small amounts of data from the ritual domain is feasible. This corresponds to the scenario of fully supervised techniques for domain adaptation discussed in the literature [6]. We experimented with different combination techniques, which are outlined in the following section.

4.2.1 Data sets

Our training data comes from two different sources. We manually annotated 408 sentences of our ritual descriptions with part of speech tags and chunks, using the Penn Treebank tagset and the CoNLL 2000 style of marking chunks [28]. As a second domain corpus we chose the Wall Street Journal, which features compatible part of speech and chunk annotations. For the extraction of chunks from the Penn Treeebank we made use of the CoNLL 2000 scripts. They were also used for the evaluation of the chunker.

We used 10-fold cross-validation to evaluate the data. In order to make sure that our test data did not include any non-ritual data, we "folded" the ritual before mixing it with the Wall Street Journal data. The significance tests are performed against a significance level of $\sigma = 0.95$ using approximate randomization [21, 22].

Table 1. Training sets for part of speech tagger and chunker

Name	Description	Sentences (one fold)	Tok./S.
WSJ	The Wall Street Journal	43,411	27.2
RIT	Ritual Descriptions	343	22.0
WSJ + RIT	Union	43,754	
WSJ + RIT ↑	oversampling RIT	86,822	
WSJ ↓ + RIT	undersampling WSJ	734	
WSJ × RIT	Combined feature space [6]	24,716	
WSJ × RIT ↑	oversampling RIT	50,785	
WSJ ↓ × RIT	undersampling WSJ	702	

Table 1 shows the different data sets and the sizes of one (average) training fold. WSJ + RIT is a simple union of the two sets. As the sizes of the two data sets differ vastly, we also experimented with equally sized corpora, by use of over- and undersampling. WSJ + RIT ↑ represents the union of the WSJ with the oversampled RIT corpus, WSJ ↓ + RIT stands for the union of the undersampled WSJ corpus with the RIT corpus.

The data set WSJ × RIT was produced by augmenting the feature space along the lines of the work in [6]. Let $\vec{v}_i =< f_1, f_2, ..., f_n >$ be the original feature vector for item i and d be a function returning an identifier for a domain. $d(0)$ is then a string representing the general domain, $d(1)$ the domain of rituals and $d(2)$ the domain of news articles. $f_k^{d(x)}$ is the same feature value as f_k, but prefixed with $d(x)$, a domain identifier. The augmented feature vector is then $\vec{v}'_i =< f_1^{d(0)}, f_2^{d(0)}, \ldots, f_n^{d(0)}, f_1^{d(i)}, f_2^{d(i)}, \ldots, f_n^{d(i)} >$, with $i = 1$ or 2. This way, each training example is annotated with a general domain feature vector and a domain specific feature vector. The learner then can learn whether to use the general domain feature set (for which it has massive training data) or the domain specific feature set (with small training data). Again, we used the same over- and undersampling techniques as before.

4.2.2 Evaluation

Part of speech tagging. Table 2 lists the results obtained with training the POS-tagger on different data sets. We use the model trained on the WSJ data set only, i.e., without any domain adaptation, as a baseline. Its performance is 94 % accuracy.

Table 2. Part of speech tagging results with different models

Training data	Accuracy
WSJ	94.01 %
RIT	95.47 %
WSJ + RIT	97.32 %
WSJ + RIT ↑	**97.59 %**
WSJ ↓ + RIT	96.97 %
WSJ × RIT	**97.19 %**
WSJ × RIT ↑	**97.40 %**

If RIT is used as (a small) training set, the POS tagger achieves a performance of 95.47 %. Training on the union of RIT and WSJ yields a significant increase in performance (97.32 %) compared to RIT. Balancing the training sets has minor, but significant influence in both directions.

Augmenting the feature space does not yield significant improvements. Neither the difference between WSJ + RIT and WSJ × RIT nor the difference between the two augmented models is significant.

Table 3. Chunking results with different models

Training data	Precision	Recall	$F_{\beta=1}$
WSJ	87.72 %	87.23 %	87.47
RIT	91.09 %	89.85 %	90.47
WSJ + RIT	90.18 %	89.44 %	89.80
WSJ + RIT ↑	91.07 %	89.62 %	90.33
WSJ ↓ + RIT	91.46 %	90.34 %	**90.89**
WSJ × RIT	88.98 %	88.15 %	88.56
WSJ × RIT ↑	91.75 %	90.24 %	**90.99**
WSJ ↓ × RIT	91.49 %	90.44 %	**90.96**

Chunking. Table 3 shows the results of the chunking models trained on the different data sets. The model trained on both the undersampled Wall Street Journal and ritual descriptions (WSJ ↓ + RIT) performed significantly better than most of the other models (90.89). The two models RIT and WSJ + RIT ↑ perform slightly lower, while not significantly different from each other. The WSJ-model achieves an F-score of only 87.47 and is thus the model with the lowest performance. Using unbalanced data (WSJ + RIT) scores significantly lower than balanced data.

The use of an augmented feature space with balanced data, as represented by data sets WSJ × RIT ↑ and WSJ ↓ × RIT, performs slightly, but not significantly, better than the best standard model. The augmented model used with an unbalanaced data set (WSJ × RIT) performs even lower than the same data set with un-augmented data (88.56).

4.3 Anaphora and coreference resolution

In order to extract continuous and consistent event chains, it is necessary to link anaphoric expressions such as pronouns (8) to their antecedents. In order to study overall performance and potential out-of-domain effects, we applied several anaphora and coreference resolution systems to the same ritual description and evaluated the labeling results.

(8) Let him give a golden coin as ritual fee [. . .].

4.3.1 Candidate systems

GuiTAR [23] and BART [33] are both modular toolkits for experimenting with different algorithms that generate entire coreference chains.

GuiTAR contains an implementation of the rule-based MARS pronoun resolution algorithm [20] and a partial implementation of an algorithm for resolving definite descriptions [34]. The latter part of GuiTAR uses the Charniak parser for preprocessing [5].

BART is a machine learning toolkit which uses a variety of features [29] to train a maximum entropy learner. In order to extract all features, the data need to be parsed or at least chunked. Additional features can be extracted from knowledge resources such as Wikipedia. In our experiment, we did not exploit BART's tuning possibilities but used the standard classifier.

In contrast to BART, JavaRAP is a rule-based anaphora resolution system that implements the Lapping & Leass algorithm for pronominal anaphora resolution [17]. It exclusively treats third person pronouns and lexical anaphors like reflexives and reciprocals and recognizes pleonastic pronouns. While BART and GuiTAR compute full coreference chains, JavaRAP only generates pairs of anaphors and antecedents. JavaRAP also uses the Charniak parser for preprocessing. Here, sentence splitting for parsing was done manually.

4.3.2 Evaluation

We evaluated these systems on a sample ritual description consisting of 295 sentences. We exclusively evaluated the resolution of personal and possessive pronouns in third person such as *it* or *him*. Anaphors which occur in direct speech are disregarded. This leaves us with 18 anaphors for evaluation.

For JavaRAP, we only evaluated anaphora-antecedent pairs. Such a pair was considered correct if the anaphoric relation to (one of potentially several) antecedents was correct. We measure an accuracy of 55.6% correctly resolved pronoun-antecedent pairs. Although this is a reasonably good outcome, the system does not build coreference chains, hence only delivers partial information.

For GuiTAR and BART, we evaluated the coreference chains which contain at least one anaphora using the scorer implemented for the SemEval-2010 coreference resolution task [25]. We measured the standard precision and recall for mention identification, and the MUC precision and recall metric for coreference resolution [35]. The MUC metric emphasizes the correctness of links between mentions within coreference chains while barely penalizing incorrect links between different chains [7]. More specifically, MUC precision is calculated by dividing the number of links in the system output that match the manual annotations by the total number of links in the system output. MUC recall is the ratio between the number of links common to the manual annotation and the system output and the total number of manually annotated links.

As a baseline, we applied the simple heuristic of resolving a pronoun to the nearest preceding noun phrase, without considering further syntactic or morphological information.

Table 4 shows the results of the mention detection task. A mention is identified as strictly correct if the system returns exactly the token of the gold standard. If a system identifies a substring of a gold mention, it is counted as partially correct. The sum of strictly and 0.5 times partially correct identified mentions is used as number of true positives.

As we can see, BART correctly identifies most of the mentions (see the low number of false negatives), however, it tends to overgenerate, with a high number of 'invented' mentions (false positives).

Table 4. Evaluation results for mention identification

Measure	Baseline	GuiTAR	BART
Total found (of 41)	36	52	60
Strictly correct	21	30	35
Partially correct	0	1	1
False positives	15	21	24
False negatives	20	10	5
Precision	58.33%	58.65%	59.16%
Recall	51.21%	74.39%	86.58%
$F_{\beta=1}$	54.54	65.59	70.29

GuiTAR both invents and identifies less mentions than BART. Both systems perfom well above the baseline system.

Table 5. Evaluation results for coreference resolution

Measure		Baseline	GuiTAR	BART
MUC	P	16.66%	50.0%	52.72%
	R	8.82%	61.76%	85.29%
	$F_{\beta=1}$	11.53	55.26	65.16

Table 5 shows precision, recall and f-measures using the MUC metric for coreference resolution. In terms of precision, BART outperforms GuiTAR by 2.72%. Comparing the recall values, BART scores more than 20% higher than GuiTAR.

Error analysis. Investigation of the analyses showed that – across all systems – the proposed coreference chains often contain correct anaphora-antecedent pairs. However, these are extended to incorrect chains, by further linking them to wrong noun phrases and pronouns as antecedents. Example 9 shows a snippet of a coreference chain computed by BART, which resolves an anaphor both correctly and incorrectly.

(9) Continue worship of the ancestors$_1$. Now at the auspicious time bring the girls$_1$ holding their$_1$ hands reciting the mantra.

Such errors also happen when gender agreement is obviously not fulfilled, as shown in example 10.

(10) The father$_2$ should touch the girl$_2$ [...] Let him$_2$ say:

Generally, both GuiTAR and BART tend to overgenerate, proposing in particular coreference chains that do not contain any anaphors. Although the obtained performance measures represent typical rates for state-of-the-art coreference resolution systems, a precision of less than 60% for the generated coreference chains is insufficient for using automatic coreference resolution on a grand scale and using its results as a basis for computing high-quality event chains. In order to obtain a utilizable coreference resolution component, we are planning to experiment with system combination techniques, such as voting and meta learning, using small amounts of annotated domain data.

5 ANNOTATION OF RITUAL DESCRIPTIONS

We use frame semantics [11] as a representation format to encode the ritual sequences such that each separable action mentioned in the ritual corpus is represented by its own frame. The actors that perform the ritual actions as well as objects, times and places mentioned are annotated as frame roles.

In a first phase, we will start with manual annotations, concentrating on developing a suitable frame inventory for the ritual domain. With an established frame inventory and an initial corpus of annotations, we will train a role semantic labeler [9] to explore automatic or semi-automatic annotation.

5.1 Adaptation of existing resources

To guarantee a consistent encoding of ritual frames, the FrameNet lexicon and ontology is used to deliver a base inventory of frames. We try to map the ritual actions to frames that are already defined in FrameNet. For this sake, verbs found in the ritual descriptions are extracted automatically from the chunked ritual descriptions. They are ordered in semantic groups and subsequently searched for in the FrameNet database. This approach has the advantage that we can make use of a well structured inventory of frames.

Coverage. According to a first estimation reported in [26], over 80% of the verbs mentioned in the ritual corpus are contained as lexical units in FrameNet. However, a closer inspection of the ritual data reveals that numerous terms are only identical at a lexical level, but occur in completely different senses. Moreover, a large number of concepts that are important in ritual descriptions are not dealt with in FrameNet. At the current state of annotation, it is difficult to give comprehensive quantitative statements about the coverage of FrameNet on ritual corpora. However, areas not (or only scarcely) covered by FrameNet include, for example, the fields of preparing and serving food.

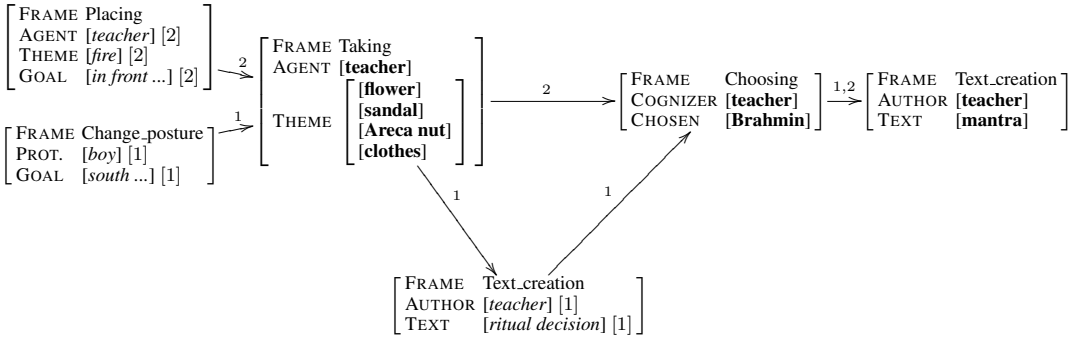
Granularity. Frequently, the frames contained in FrameNet represent concepts that are too abstract for the annotation of rituals. In these cases, FrameNet groups several lexical units into one frame that would not correspond to a single concept in a genuine ritual frame ontology. Examples are the verbs “to cover”, “to anoint” and “to fill”. These verbs are assigned to a single frame *Filling* in FrameNet because they express the same idea of “filling containers and covering areas with some thing, things or substance”. Although we use frames to generalize from the literal level of ritual descriptions, annotating “to fill” and “to anoint” by a single frame in a ritual context would certainly lead to over-generalization and, therefore, to a clear loss of information. New frames have to be designed in such cases. For the example under consideration, we decided to further specify the frame Filling with three more specialized new frames: *Filling_container* (filling a container), *Besmearing_surface* (covering a surface with a liquid) and *Wrapping_object* (wrapping an object with another object).

On the other hand, the granularity of FrameNet frames can also be higher than needed. This case occurs, for instance, in the area of legal concepts, which are covered in great detail by FrameNet. Such cases are easier to resolve than those resulting from coarse-grainedness of frames discussed above, due to the FrameNet inheritance hierarchy. That is, we can use predefined, more abstract frames from higher levels in the hierarchy.

The existence of such selected semantic fields that are covered in FrameNet in great detail clearly demonstrates that it has been successful in modeling specific domains. Thus, for the present project, domain adaptation will consist in modeling finer-grained frame structures for semantic fields that are relevant for the ritual domain.

Annotation and automation. The mapping from verbs to frames is stored in an index and used to assign frames automatically to the

Figure 1. A schematic representation of a common subsequence in two different rituals; the indices indicate the number of the example.



verbs in the preprocessed ritual descriptions. Currently, we have defined 116 of such predefined assignment rules. Applying them to two ritual descriptions yielded coverage rates of 35.2% (479 of 1361 verbal units) for a modern ethnographic report and 82.5% (254 of 308 verbal units) for the translation of an indigenous manual (cf. 3.1), respectively. A closer inspection of the latter reveals that three frames contribute 65.3% to the high coverage. This is caused by the rather monotonous character of this text whose main part consists of the repeated invocation of Hindu deities ("Salutation to god ..."; mapped to a newly designed frame in 88 instances) and describes the recitation of mantras (mapped to FrameNet *Text_creation*, 91 instances) and the offering of ritual stuff to the participants and deities (FrameNet *Giving*, 22 instances).

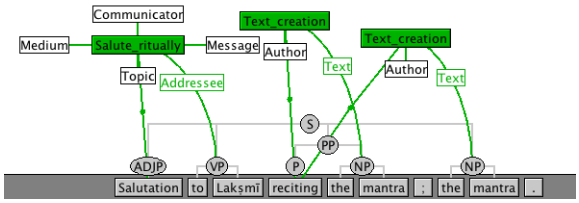


Figure 2. Annotated sentence *Salutation to Lakṣmī reciting the dyāṃ mā lekhīr; the śrīś ca te.*

The Salsa tool [8] (Figure 2) is used to manually correct the automatic annotations and to assign frame-semantic roles to syntactic constituents. This frame semantic information is stored as a separate layer along with the source text and the linguistic annotation. When the corpus text or the linguistic preprocessing layers are updated, this layering mechanism makes it possible to reassign the frame semantic annotation, thus avoiding manual re-annotation.

5.2 Detecting ritual structure

As a proof of concept for the types of analyses we can offer to ritual scientists on the basis of these semantic annotations, we constructed representations for a number of close variations of rituals, two of which are shown below with assigned frames shown as subscripts.

1. "... the boy sits down$_{\text{CHANGE_POSTURE}}$ south of the teacher. (The teacher) takes$_{\text{TAKING}}$ flowers, sandal, Areca nut and clothes and declares$_{\text{TEXT_CREATION}}$ the ritual decision to select$_{\text{CHOOSING}}$ the Brahmin by saying$_{\text{TEXT_CREATION}}$ the mantra ..."
2. " ... (the teacher) places$_{\text{PLACING}}$ (fire in a vessel of bell metal) in front of himself. Having taken$_{\text{TAKING}}$ flowers, sandal, Areca nut, clothing etc. he should select$_{\text{CHOOSING}}$ a Brahmin. The Brahmin is selected with$_{\text{TEXT_CREATION}}$ the mantra ..."

We extracted the event sequences from each description, one starting with PLACING, one with CHANGE_POSTURE. Figure 1 shows a partial semantic representation for the above excerpts. It illustrates one way in which we plan to extract and visualize common subsequences in rituals. The sequences share the frames TAKING, CHOOSING and TEXT_CREATION. Elements occurring in both sequences are printed in bold.

6 FUTURE WORK AND CONCLUSIONS

6.1 Future work

As we have seen, anaphora resolution is currently an unsolved issue. We intend to perform a detailed error analysis of the available systems and to identify strategies and methods that can yield reasonable performance with respect to the overall task.

Several other steps in the preprocessing chain that have not been discussed in this paper need to be addressed in the future. Word sense as well as named entity annotations are needed as a basis for semantic annotation and the structural analysis of rituals. As we established in a pre-study, many ritual specific concepts are not included in sense inventories such as WordNet. Also, named entities occurring in ritual descriptions can often not be classified into the standard classes or do not appear in gazetteer lists. Thus, we expect that both word sense disambiguation and named entity recognition systems and resources need to be adapted to the ritual domain.

Using the types of annotations discussed in this paper, we will create structured and normalized semantic representations for ritual descriptions that are linked to an ontology comprising general-semantic and ritual-specific concepts and relations. This allows us to offer querying functionalities for ritual researchers, so that they can test and validate their hypotheses against a corpus of structurally analyzed ritual descriptions. A well-defined and populated ontology can also be used to automatically identify event sequences in the data.

Sequence analysis and the automatic detection of structure in rituals are the second focus of our future research. As soon as enough data has been encoded in the scheme described in sections 4 and 5, we plan to develop computational methods that support ritual researchers in finding constant patterns and variations in the ritual descriptions. Methods that will be adapted for this purpose include modeling of selectional preferences, as well as algorithms for detecting frequent item sets and statistical tests of significance.

6.2 Conclusions

In this paper, we presented a detailed investigation of the performance of standard NLP tools and resources for the computational linguistic analysis of ritual descriptions. As standard "out of the box" tools perform poorly and lexical resources are lacking coverage and the appropriate granularity, the adaptation of tools and resources to different domains emerges as an important focus of our work. However, we have not only established that standard NLP tools behave poorly on our domain, we also have shown that we can improve the results significantly with rather small effort. This finding supports our basic tenet, that it is possible to make use of computational linguistics methods for the semantic and quantitative analysis of ritual texts. Further work will have to establish whether the representations we compute will allow us to help ritual researchers establish novel insights on the structure(s) of rituals.

Our work also explores to which degree methods of computational linguistics can be adapted to the needs of the Humanities. By using a rarely applied combination of computational and traditional scholarship, we are optimistic to achieve results that extend the knowledge in the field of ritual research to a considerable degree. Moreover, we hope to open up new, more formal data-oriented ways for research in the Humanities.

ACKNOWLEDGEMENTS

This research has been funded by the German Research Foundation (DFG) and is part of the collaborative research center on ritual dynamics (Sonderforschungsbereich SFB-619, Ritualdynamik).

REFERENCES

[1] A. Burchardt, A. Frank, and M. Pinkal, 'Building Text Meaning Representations from Contextually Related Frames – A Case Study', in *Proceedings of IWCS*, (2005).

[2] A. Burchardt, S. Pado, D. Spohr, A. Frank, and U. Heid, 'Constructing integrated corpus and lexicon models for multi-layer annotations in owl dl', *Linguistic Issues in Language Technology*, **1**(1), 1–33, (2008).

[3] N. Chambers and D. Jurafsky, 'Unsupervised learning of narrative event chains', in *Proceedings of ACL: HLT*, pp. 789–797, (2008).

[4] N. Chambers and D. Jurafsky, 'Unsupervised learning of narrative schemas and their participants', in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 602–610, (2009).

[5] E. Charniak, 'A Maximum-Entropy-Inspired Parser', in *Proceedings of NAACL*, (2000).

[6] H. Daumé III, 'Frustratingly easy domain adaptation', in *Proceedings of ACL*, pp. 256–263, (2007).

[7] P. Denis, *New learning models for robust reference resolution*, Ph.D. dissertation, Austin, TX, USA, 2007. Adviser-Baldridge, Jason M. and Adviser-Asher, Nicholas M.

[8] K. Erk, A. Kowalski, and S. Padó, 'The SALSA Annotation Tool', in *Proceedings of the Workshop on Prospects and Advances in the Syntax/Semantics Interface*, (2003).

[9] K. Erk and S. Padó, 'Shalmaneser – a Toolchain for Shallow Semantic Parsing', in *Proceedings of LREC*, (2006).

[10] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[11] C. J. Fillmore, C. R. Johnson, and M. R. Petruck, 'Background to FrameNet', *International Journal of Lexicography*, **16**(3), 235–250, (2003).

[12] J. R. Finkel and C. D. Manning, 'Hierarchical Bayesian Domain Adaptation', in *Proceedings of HLT-NAACL*, pp. 602–610, (2009).

[13] O. Hellwig, 'A chronometric approach to Indian alchemical literature', *Literary and Linguistic Computing*, **24**(4), 373–383, (2009).

[14] K. Hyland, 'Hedging in academic writing and eap textbooks', *English for Specific Purposes*, **13**(3), 239–256, (1994).

[15] J. Jiang and C. Zhai, 'Instance Weighting for Domain Adaptation in NLP', in *Proceedings of ACL*, pp. 264–271, (2007).

[16] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer, 'A Large-Scale Classification of English Verbs', *Journal of Language Resources and Evaluation*, **42**(1), 21–40, (2008).

[17] S. Lappin and H. J. Leass, 'An algorithm for pronominal anaphora resolution', *Computational Linguistics*, **20**(4), 535–561, (1994).

[18] M. Light, X. Y. Qiu, and P. Srinivasan, 'The Language of Bioscience: Facts, Speculations, and Statements In Between', in *Proceedings of HLT-NAACL Workshop: BioLINK*, eds., L. Hirschman and J. Pustejovsky, pp. 17–24, (2004).

[19] B. Medlock and T. Briscoe, 'Weakly Supervised Learning for Hedge Classification in Scientific Literature', in *Proceedings of ACL*, pp. 992–999, (2007).

[20] R. Mitkov, *Anaphora Resolution*, Longman, 2002.

[21] E. W. Noreen, *Computer-Intensive Methods for Testing Hypotheses*, John Wiley & Sons, 1989.

[22] S. Padó, *User's guide to* `sigf`*: Significance testing by approximate randomisation*, 2006.

[23] M. Poesio and M. A. Kabadjov, 'A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation', in *Proceedings of LREC*, (2004).

[24] S. Pradhan, W. Ward, and J. H. Martin, 'Towards Robust Semantic Role Labeling', *Computational Linguistics, Special Issue on Semantic Role Labeling*, **34**(2), 289–310, (2008).

[25] M. Recasens, T. Martí, M. Taulé, L. Màrquez, and E. Sapena, 'Semeval-2010 task 1: Coreference resolution in multiple languages', in *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future*, pp. 70–75, (2009).

[26] N. Reiter, O. Hellwig, A. Mishra, A. Frank, and J. Burkhardt, 'Using NLP methods for the Analysis of Rituals', in *Proceedings of LREC*, (2010).

[27] J. Ruppenhofer, C. Sporleder, R. Morante, C. Baker, and M. Palmer, 'Semeval-2010 task 10: Linking events and their participants in discourse', in *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pp. 106–111, (2009).

[28] E. F. T. K. Sang and S. Buchholz, 'Introduction to the CoNLL-2000 Shared Task: Chunking', in *Proceedings of CoNLL-2000 and LLL-2000*, (2000).

[29] W. M. Soon, D. C. Y. Lim, and H. T. Ng, 'A machine learning approach to coreference resolution of noun phrases', *Computational Linguistics*, **27**(4), 521–544, (December 2001).

[30] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik, 'The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts', in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 38–45, (2008).

[31] K. Toutanova, D. Klein, C. Manning, and Y. Singer, 'Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network', in *Proceedings of HLT-NAACL*, pp. 252–259, (2003).

[32] M. Verhagen and J. Pustejovsky, 'Temporal processing with the TARSQI toolkit', in *Coling 2008: Companion volume: Demonstrations*, pp. 189–192, Manchester, UK, (August 2008). Coling 2008 Organizing Committee.

[33] Y. Versley, S. P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, and A. Moschitti, 'Bart: A modular toolkit for coreference resolution', in *Proceedings of the ACL: HLT Demo Session*, pp. 9–12, (2008).

[34] R. Vieira and M. Poesio, 'An empirically-based system for processing definite descriptions', *Computational Linguistics*, **26**(4), (2000).

[35] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, 'A model-theoretic coreference scoring scheme', in *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pp. 45–52, Morristown, NJ, USA, (1995).